ABSTRACT

Title of dissertation: Data Representation for Learning and Information Fusion in Bioinformatics

Vinodh N. Rajapakse, Doctor of Philosophy, 2013

Dissertation directed by: Professor Wojciech Czaja
Department of Mathematics

This thesis deals with the rigorous application of nonlinear dimension reduction and data organization techniques to biomedical data analysis. The Laplacian Eigenmaps algorithm is representative of these methods and has been widely applied in manifold learning and related areas. While their asymptotic manifold recovery behavior has been well-characterized, the clustering properties of Laplacian embeddings with finite data are largely motivated by heuristic arguments. We develop a precise bound, characterizing cluster structure preservation under Laplacian embeddings. From this foundation, we introduce flexible and mathematically well-founded approaches for information fusion and feature representation. These methods are applied to three substantial case studies in bioinformatics, illustrating their capacity to extract scientifically valuable information from complex data.

DATA REPRESENTATION FOR LEARNING AND INFORMATION FUSION IN BIOINFORMATICS

by

Vinodh Nalin Rajapakse

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Wojciech Czaja, Chair/Advisor
Professor John J. Benedetto
Professor Michael P. Cummings
Professor Eric S. Haag
Professor Kasso A. Okoudjou
Dr. Yves G. Pommier

Dedication

To the cherished memory of my grandparents:

Dionicious and Mildred Gunawardena

and

Simeon and Sirimawathie Rajapakse.

Acknowledgments

I must begin by extending my heartfelt gratitude to my advisor, Professor Wojciech Czaja. There is a lot to advertise about Wojtek - notably the depth and strength of his mathematical abilities, as well as his broad curiosity and genuine interest in finding novel mathematics to address application areas. But in the end, what I will remember the most is simply that, beyond my family, no one has exceeded Wojtek's incredible generosity and enduring concern for my best interests. Anyone who knows me and my amazingly supportive family will understand how much I mean by this. I will always be grateful for the opportunities that Wojtek has opened for me. Among the first of these was affiliation with the Norbert Wiener Center, and the great privilege of meeting and learning from Professor John Benedetto. Recalling my undergraduate studies in biology, I first wondered if I even belonged in the same lecture room as John and his talented students. But as many can attest, John is such a gifted teacher, and beyond this, such a warm and wonderful person, that I soon relaxed and started learning.

There are so many others to thank just at the University of Maryland, but I want to start with my Norbert Wiener Center colleagues. I will always remember Professor Dennis Healy, for his kindness and his inspiring, enthusiastic brilliance. I owe a great deal to Ioannis Konstantinidis, who really launched my graduate studies by inviting me to join the NWC's MAIT program. Ioanni also introduced me to Martin Ehler, who I was incredibly fortunate to work with and learn from. Martin also very helpfully led me to Wojtek! I must also thank Professor Radu Balan, for

being so supportive of my work, and Professor Kasso Okoudjou, who very kindly joined my dissertation committee, and then took the time to discuss further research directions. The many students and post-docs associated with the NWC have clearly been a great source of support, and I want to especially thank Julia Dobrosotskaya, Tom McCullough, and Nate Strawn.

Many others at Maryland enabled my graduate studies: Professors Paul Smith, Matei Machedon, and Robert Warner all taught courses that greatly helped me on my journey from biology to applied math. I am very grateful to Professor Konstantina Trivisa and Mrs. Alverda McCoy, for graciously admitting and supporting me through the AMSC program. Last, and far from least at UMD, I must thank Professors Eric Haag and Michael Cummings, two biologists who kindly joined my dissertation committee, and also provided me with very helpful advice on the substantial application-focused elements of my work.

Through my graduate studies, I was fortunate to have the chance to do research at the National Institutes of Health. This work began with Drs. Robert Bonner and Barry Zeeberg, who taught me a great deal about bridging the quantitative and biological sciences. Through the latter part of my graduate studies, I have had the wonderful opportunity to work at the National Cancer Institute's Laboratory of Molecular Pharmacology, under the generous mentorship of Dr. Yves Pommier and Mr. William Reinhold. This has been one of the most rewarding experiences of my career, and I am incredibly grateful for their support, together with that of my LMP colleagues: Augustin Luna, Sudhir Varma, Fabricio Sousa and Margot Sunshine.

In 2003, when I first moved to Maryland to take a position at the National Institutes of Health, I could not have imagined that I would ultimately pursue graduate studies in a mathematics department. I must thank two authoritative people who helped me to recognize that this was a viable and worthwhile course: my biotechnology industry mentor, Dr. Scott Markel, and my dear friend Ray Jayawardhana.

Finally, I want to extend the greatest thanks to my incredible, inspiring family members for their boundless love and support, in every sense of the word. Nothing I have achieved would have been possible without them - my parents Vijitha and Nelunika Rajapakse, my uncles Shanti and Ramesh Gunawardena, my sister Mimi, my cousins Delani, Devaka, and Andy, as well as John and Maggie.

Table of Contents

List of Figures

Chapter 1

Introduction

Fundamental advances in molecular biology have driven the development of increasingly sophisticated molecular profiling technologies, which have in turn opened new scientific horizons. A notable product of this virtuous cycle has been the explosive growth of biological data. Gene expression microarrays and next-generation sequencing technologies allow the measurement of thousands of genes in collections of biological samples. The complex data sets that result can be regarded as coarse snapshots of biomolecular network states and their aggregate output. These underlying networks are complex and dynamic, but at the same time, they clearly possess considerable structure. A basic challenge is to recover elements of this structure in ways that can yield new scientific insights for experimental development.

The starting point in this effort is naturally the data generated by profiling technologies, which are typically complex, noisy, and high-dimensional. The approach developed in this thesis is to consider *data representations* that can organize this sort of complex, high-dimensional data in a manner that reveals fundamental structure. A widely-known example of a data representation approach is Principal Component Analysis (PCA) [61], which can reorganize and simplify data that are concentrated on a linear subspace. In this work, we examine, theoretically and empirically, the properties of the Laplacian Eigenmaps algorithm, [4] which can be

viewed as a nonlinear analogue of PCA.

Laplacian Eigenmaps (LE) is often described as a nonlinear dimension reduction technique, and theoretical results have indeed established its capacity to recover essential features of manifold-structured data [5]. In view of the highly nonlinear structure and dynamics of biological networks, this flexibility is notably appealing. But, with biological data, broader structural attributes are often of immediate interest. Among these, cluster structure is perhaps the most widely considered. Across diverse data types, clustering allows information to be propagated from limited numbers of known entities that are co-organized amidst poorly understood ones.

Attributes of Laplacian Eigenmaps, and closely related methods such as Diffusion Maps [18], suggest an appealing capacity to resolve a very broad range of cluster structure. This notably includes clusters with complex geometries, which can arise in high-dimensional data derived from complex systems. These clustering properties have largely been suggested by informal arguments. In this thesis, we develop a precise characterization of cluster structure preservation under Laplacian-based embeddings. From this foundation, we show how Laplacian-based data representation methods can be applied to flexibly combine information from different data sources.

We start in Chapter 2 with an overview of dimension reduction and data representation techniques, focusing on Laplacian Eigenmaps and the related Diffusion Maps method. In Chapter 3, we examine the clustering properties of Laplacian-based data embeddings. Building on the work of Hunter and Strohmer [43], we develop a result characterizing the effect of a Laplacian matrix perturbation on the Laplacian-based data embedding. We show that for n well-separated clusters of

arbitrary geometry, Laplacian-based representations map all points to orthogonally separated, cluster-specific points that are further organized according to the intra-cluster connectivity. Applying the developed perturbation result, we show that less separated clusters are organized around the separated-case, cluster-specific points with a bound governed by fundamental cluster structure attributes, such as the internal coherence of the clusters, as well as their inter-cluster connectivity.

In Chapter 4, we introduce methods for two types of information fusion. In the first 'multiview fusion' case, we have multiple data sets presenting different kinds of measurements for a fixed collection of elements. We show that multiple kernels describing data set-specific relationships can be flexibly combined within a direct extension of the Laplacian-based data representation framework. A more general sort of data fusion entails combining data sets recording observations of distinct and non-overlapping sets of elements. Certain relationships are known or inferred to exist between subsets of elements in the different data sets. The aim is to construct a joint, heterogeneous data embedding that 'aligns' the data sets with respect to these related elements, while preserving their respective structures. Building on a recently developed generalization of the Diffusion Maps framework by Coifman and Hirn [17], we introduce a novel, frame-based algorithm for constructing a joint embedding of two distinct but related data sets.

With the mathematical foundation developed in Chapters 2 - 4, we present in Chapter 5 three substantial case studies in biological data analysis. In the first, we apply Laplacian and Schrödinger Eigenmaps to analyze a microarray gene expression data set from a study of vertebrate eye development. We compare clusters developed

using the Laplacian-based methods with ones derived from the original data and PCA-processed data, and show greater biological specificity in the LE-based clusters. This work was published in [25].

In the second case study, we apply the multi-kernel, Laplacian-based information fusion methods introduced in Chapter 4 to organize genes with respect to a combined measure of co-expression and proximity along a chromosome. The aim is to predict features of chromosomal domain organization that may be related to coordinated gene expression. A network of putative expression-related inter-chromosomal interactions is constructed, and the results are assessed statistically, and with respect to measured chromosomal interactions. This work was published in [63].

In the third case study, we apply Laplacian Eigenmaps to organize a large database of drug compound chemoactivity profiles and identify coherent clusters of compounds sharing similar response profiles over the NCI-60 cancer cell lines. The clusters are additionally organized in a network based on their relative similarity. This drug cluster network is shown to be highly concordant with the existing understanding of compound class relationships, grouping known mechanism of action drugs into coherent clusters, while revealing groups of novel compounds sharing similar response profiles. The drug cluster network is integrated with a gene co-expression network to identify sets of co-expression modules that are potentially implicated in drug responses. We additionally present initial results from computational experiments with the joint embedding algorithm presented in Chapter 4, applied to co-organize genes and drug compounds directly.

Chapter 2

Background

2.1 Dimension Reduction and Data Organization

Large, high-dimensional data sets are increasingly encountered in many areas of science, with the ability to collect data often outstripping the means to effectively analyze it. An encouraging prospect is that complex data sets frequently possess elements of organizing structure. For example, numerous sensors may redundantly record attributes of a process that is fundamentally driven by a much smaller number of parameters. Some common challenges include, among others, nonlinear interaction of parameters or corruption by noise. These features are likely to limit established analysis techniques, such as Principal Components Analysis (PCA) [61], that presume a relatively simpler, linear data structure. To address this situation, a number of nonlinear data organization approaches have been developed, such as Kernel PCA [70], ISOMAP [78], Locally Linear Embedding (LLE) [69], Hessian LLE [23], Laplacian (Schrödinger) Eigenmaps [4, 5, 20], and Diffusion Maps [18, 19]. These methods all aim to reduce the apparent complexity of data sets by mapping their points to a space of lower dimension, while preserving important elements of the original structure or geometry. By geometry, we mean the intrinsic relationships between data points, e.g., which elements are 'connected'. Connection here is defined in some suitable, application-specific manner, as are the appropri-

ately preserved structural features of the data. In this chapter, we will present some background on Laplacian Eigenmaps and several closely related *normalized output* algorithms. These methods share a common approach, presented in [29], for recovering the low-dimensional structure of an input data set. They first identify and represent the local neighborhood structure of the data. Then, an embedding is constructed by solving a particular convex optimization problem subject to certain normalization constraints. The latter constraints impose a degree of local structure preservation which can, in particular, enable resolution of a broad range of cluster structure. These flexible clustering properties support and enhance a range of learning techniques, motivating our focus on Laplacian-based embeddings. Relationships between the presented approaches will be discussed, followed by some examples that illustrate their properties.

2.2 Preliminaries and Notation

To unify the presentation through this chapter and the whole thesis, we review some foundational items and establish the following notation.

- N is the number of points in the input data set.
- D is the dimension of the input data, and d is the dimension of the output data.
- The high-dimensional input data points will be specified as $x_1, ..., x_N \in \mathbb{R}^D$, and these are organized into the $N \times D$ matrix X, with the i-th row representing x_i.

- The data representation output by a particular algorithm will be specified as $y_1, ..., y_N \in \mathbb{R}^d$, with the output points organized as the rows of the $N \times d$ matrix Y.

The normalized output algorithms model the data as a graph, with points $x_1, ..., x_N \in \mathbb{R}^D$ identified with vertices, and undirected edges representing relationships between points. We use the standard notation $G = (V, E)$ to denote an undirected graph with vertex set V and edge set E. Given a data set X, there are several approaches for constructing a data graph based on pairwise similarities s_{ij} or pairwise distances d_{ij} between points x_i and x_j [82].

- The *ϵ-neighborhood graph* places an edge between all points separated by a distance less than ϵ. Since ϵ is chosen so that distances between connected points are of the same scale, the edges are typically not weighted.

- The *k-nearest neighbor graph* places an edge between the vertices identified with points x_i and x_j if x_i is one of the k nearest neighbors of x_j or vice versa. An alternative approach for obtaining a symmetric neighborhood relationship is to construct the *mutual k-nearest neighbor graph*. Here points x_i and x_j are connected only if x_i is among the k nearest neighbors of x_j *and* x_j is one of the k nearest neighbors of x_i. For both of these neighborhood graphs, edges can be weighted or unweighted.

- The *fully connected graph* relates all points, with edge weights given by a similarity function that effectively models local neighborhood relationships. A

popular choice is the Gaussian kernel $e^{\frac{-\|x_i - x_j\|^2}{\sigma}}$, $\sigma > 0$. Here the kernel bandwidth parameter σ controls the neighborhood extent in a manner comparable to the parameter ϵ in the ϵ-neighborhood graph.

The normalized output algorithms derive data representations from the eigenvectors of matrices constructed from the data graph. These matrices are real-valued and symmetric, so the eigendecomposition is guaranteed by the following result [51, 75].

Theorem 1 (Spectral Theorem)**.** *Let A be any real, symmetric $n \times n$ matrix. Then:*

1. *A has n real eigenvalues $\lambda_1, ..., \lambda_n$ (not necessarily distinct).*
2. *A has a set of n eigenvectors $u_1, ...u_n$ that form an orthonormal basis for $\mathbb{R}^n$, that is $u_i^T u_j = \delta_{ij}$, for all i, j.*

We thus have the spectral decomposition

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T = U \Lambda U^T,$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$ and $U = [u_1, ...u_n]$ is an orthogonal matrix with the eigenvectors of A along its columns.

We accordingly assume that the eigenvalue-eigenvector pairs (λ_i, v_i) are ordered with respect to the magnitude of the eigenvalues. If we have $\lambda_1 \leq \lambda_2, ..., \lambda_N$, we refer to $v_1, ..., v_d$ as the *bottom d* eigenvectors, and $v_{N-d+1}, ..., v_N$ as the *top d* eigenvectors. We finally recall the important class of positive semidefinite matrices, together with a useful way of characterizing them in terms of their eigenvalues.

Definition 1 (Positive Semidefinite Matrices)**.** *A real, symmetric $n \times n$ matrix A is positive semidefinite (denoted $A \succeq 0$) if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$. It is positive definite (denoted $A \succ 0$) if $x^T Ax > 0$ for all nonzero $x \in \mathbb{R}^n$.*

Theorem 2. *Let A be a real, symmetric $n \times n$ matrix. Then:*
A is positive semidefinite (positive definite) if and only if $\lambda_i \geq 0$ ($\lambda_i > 0$) for all eigenvalues λ_i, $i = 1, ..., n$.

2.3 Principal Components Analysis

Principal Components Analysis (PCA) is one of the best-known algorithms for dimension reduction and data representation. It was first described by Pearson in 1901 as method for computing "lines and planes of closest fit to systems of points in space" [61], and was further extended by Hotelling, who applied it to psychometry [42]. In the context of stochastic processes, PCA was independently developed by Karhunen and Loeve [45, 58], and is sometimes described as the Karhunen-Loeve Transform. For a given data set X, PCA identifies the directions which capture the largest amount of variation in the data. If we assume, without loss of generality, that the data is mean-centered, i.e., the mean of the data points is subtracted from

each point, the single direction of maximum variation can be computed as

$$\begin{aligned}
\arg\max_{\|v\|=1} \mathrm{Var}(Xv) &= \arg\max_{\|v\|=1} \mathbb{E}(Xv)^2 \\
&= \arg\max_{\|v\|=1} \sum_{i=1}^{N} (x_i^T v)^2 \\
&= \arg\max_{\|v\|=1} \sum_{i=1}^{N} (v^T x_i)(x_i^T v) \\
&= \arg\max_{\|v\|=1} v^T \left(\sum_{i=1}^{N} (x_i x_i^T) \right) v \\
&= \arg\max_{\|v\|=1} v^T X^T X v.
\end{aligned}$$

The last line above is maximized by setting v to be the top eigenvector of the data *covariance matrix* $X^T X$. More generally, the projection of the data onto the d orthogonal vectors associated with the greatest variation is given by XV, where V is the $D \times d$ matrix constructed from the top eigenvectors of $X^T X$. The particular eigenvalues specify the variance of the data projected along the respective eigenvectors. If there are directions along which the data varies only minimally, dimension reduction can be achieved by setting $d < D$, effectively dropping the latter directions, which may reflect noise, etc. in the data. PCA is optimal for data that lies on or very near a linear subspace. For data derived from even a simple non-linear manifold, however, it can yield poor data representations, as illustrated by the example in Section 2.8.1 of this chapter. The non-linear techniques that follow in Sections 2.4, 2.5, and 2.7 can be seen as analogues of PCA that seek to faithfully represent more complex, non-linear data.

2.4 Laplacian Eigenmaps

Laplacian Eigenmaps (LE) was described by Belkin and Niyogi in [4], and further developed by the same authors in [5]. With LE, we assume that our data set consists of points $x_1, ..., x_N$, drawn from a d-dimensional manifold in $\mathbb{R}^D$, and we assume that $d << D$. More generally, we may assume that the data is sampled from a distribution with support concentrated on a d-dimensional manifold. We obtain a manifold structure preserving low dimensional representation $y_1, ..., y_N \subset \mathbb{R}^d$ in three steps:

1. Construct Data Adjacency Matrix W: For $k \in \mathbb{N}$, put an edge between elements i and j if x_i is among the k nearest neighbors of x_j or vice versa. Weight connected edges using $W_{ij} = e^{\frac{-\|x_i - x_j\|^2}{\sigma}}$, $\sigma > 0$.

2. Construct Laplacian Matrix L: Set $D_{ii} = \sum_{j=1}^{N} W_{ij}$, and let $L = D - W$.

3. Compute Eigenmaps: Solve $Lx = \lambda Dx$. Let $f_0, f_1, ..., f_d$ be the eigenvectors corresponding to the first $d + 1$ eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_d$. Discard f_0 and embed in d-dimensional space using the map $x_i \rightarrow y_i = (f_1(i), f_2(i), ..., f_d(i))$. For a connected data graph, one can verify that f_0 is the constant one vector $\mathbb{1}$. See [15, 82] for further details.

The overall approach is motivated by the fact that the graph-based Laplacian matrix L can be seen as a discrete analogue of the Laplace-Beltrami operator $\mathcal{L}$ on the underlying manifold. In particular, Belkin and Niyogi show that the Laplacian constructed from the adjacency graph of data uniformly sampled from a compact

submanifold $\mathcal{M}$ of $\mathbb{R}^D$ converges to the appropriately normalized Laplace-Beltrami operator on $\mathcal{M}$, see [5]. The eigenmaps of the latter operator provide an optimal embedding of the manifold into a space of reduced dimension. To recognize this, suppose that $\mathcal{M}$ is a smooth, compact, d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. We would like to find a map from the manifold $\mathcal{M}$ to the real line $\mathbb{R}$ that preserves local neighborhood structure, mapping nearby points on the manifold to nearby points on the line. Let $f : \mathcal{M} \to \mathbb{R}$ denote the desired map. If we suppose that f is twice differentiable, it is not hard to show that for neighboring points $x, z \in \mathcal{M}$,

$$|f(z) - f(x)| \leq \|\nabla f(x)\| \|z - x\| + o(\|z - x\|).$$

With $\|\nabla f(x)\|$ providing an estimate of how far apart f maps nearby points, we see that the optimal locality preserving map is given by

$$\underset{\|f\|_{L^2(\mathcal{M})}=1}{\arg\min} \int_{\mathcal{M}} \|\nabla f(x)\|^2 = \underset{\|f\|_{L^2(\mathcal{M})}=1}{\arg\min} \int_{\mathcal{M}} \mathcal{L}(f) f.$$

In the above, the equality of the minimized integrals follows from the fact that $\mathcal{L} \equiv -\operatorname{div} \nabla(f)$ and from the Stokes' Theorem. Consequently, we have that the desired map f must be an eigenfunction of $\mathcal{L}$, which is positive semidefinite and has a discrete spectrum (since $\mathcal{M}$ is assumed to be compact [67]). If the eigenfunctions of $\mathcal{L}$ are ordered with respect to the eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq ...$, we note that f_0 can be excluded as it is the constant function mapping the entire manifold to a single point [4, 5]. The desired locality preserving map is thus f_1, and more generally, for an optimal d-dimensional embedding, we have $x \to (f_1(x), ..., f_d(x))$.

Since the graph-based Laplacian converges to the manifold-based Laplace-Beltrami operator, its associated data mappings progressively inherit the corresponding manifold recovery guarantees [4, 5]. These ideas motivate a more concrete understanding of LE in the finite data settting. Let the $N \times d$ matrix $Y = (y_1, ..., y_N)^T$ denote the low-dimensional representation of our data set X. The eigenvalue problem $Lx = \lambda Dx$ can be shown to solve the following minimization:

$$\underset{Y^T DY = I}{\arg\min} \operatorname{trace}(Y^T LY) = \underset{Y^T DY = I}{\arg\min} \frac{1}{2} \sum_{i,j=1}^{N} \|y_i - y_j\|^2 W_{i,j}. \tag{2.1}$$

Note that the first term on the right forces neighboring points in the original data space, i.e., with large W_{ij}, to be mapped close to one another. As such, LE acts to organize data with respect to local features, which allows natural cluster structure to be better revealed. These cluster preservation properties will be further motivated and developed in Chapter 3. For review and development of Laplacian Eigenmaps, together with presentation of numerical experiments with hyperspectral imaging data, see also [27, 32, 38, 84].

2.5 Schrödinger Eigenmaps

The Schrödinger Eigenmaps procedure, introduced by Czaja and Ehler in [20], builds on the strengths of Laplacian Eigenmaps. The Laplace Equation $\Delta\varphi = 0$ can be extended to the time-independent Schrödinger Equation by adding a potential term $v(x)$ to the Laplace operator:

$$\mathcal{E}\Psi(x) = \Delta\Psi(x) + v(x)\Psi(x). \tag{2.2}$$

The discrete analogue of the resulting Schrödinger operator $\mathcal{E} = \Delta + v$ is the matrix $E = L + V$, where V is a nonnegative diagonal matrix. It can be shown that the matrix $L + \alpha V$ can be applied in place of the matrix L in the framework presented above, with the parameter $\alpha > 0$ determining the strength of the potential [20, 32]. Let the $N \times d$ matrix $Y = (y_1, ..., y_N)^T$ denote the resulting low-dimensional representation of our data set. The modified eigenvalue problem $(L + \alpha V)x = \lambda Dx$ can be shown to solve the following minimization [20, 32]:

$$\arg\min_{Y^T DY = I} \operatorname{trace}(Y^T (L + \alpha V) Y) = \arg\min_{Y^T DY = I} \frac{1}{2} \sum_{i,j=1}^{N} \|y_i - y_j\|^2 W_{i,j} + \alpha \sum_{i=1}^{N} V(i) \|y_i\|^2. \quad (2.3)$$

Note that the first term on the right forces neighboring points in the original data space, i.e., with large W_{ij}, to be mapped close to one another. The second term, however, penalizes points with large values of $V(i)$. In the case of a binary-valued potential, the minimization would tend to push points with corresponding potential values of 1 together and toward zero. More general potentials (including non-diagonal ones) can be constructed and effectively used for introducing prior knowledge, labeling sets of points that ought to cluster together (or be separated) based on external evidence [25]. The behavior is nuanced because the minimization must balance information expressed through the potential with the data-dependent connectivity structure captured by the weights W_{ij}.

2.6 Basic Properties of Graph Laplacians

In this Section, we summarize for future reference some basic properties of graph Laplacians, following elements of the notation and presentation in [82]. Sev-

eral related graph Laplacians appear in the literature [15, 82]. The one presented in association with Laplacian Eigenmaps, $L = D - W$, is often described as the *unnormalized* graph Laplacian [82]. A straightforward computation shows that for every vector $f \in \mathbb{R}^N$, we have:

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{N} W_{ij}(f_i - f_j)^2,$$

where W_{ij} is an element of the nonnegative-valued, symmetric weight matrix W. Thus L is symmetric and positive semi-definite, with a minimal eigenvalue of 0, associated with the constant one vector $\mathbb{1}$.

Two *normalized* graph Laplacians are defined in [82]:

$$L_{rw} = D^{-1}L \quad \text{and} \quad L_{sym} = D^{-1/2}LD^{-1/2}.$$

L_{rw} is closely related to a random walk on the nodes of the graph, and from $L_{sym} = D^{1/2}L_{rw}D^{-1/2}$, we see that L_{sym} is a symmetric matrix similar to L_{rw}. We accordingly have:

$$Lx = \lambda Dx \;\Leftrightarrow\; L_{rw}x = \lambda x \;\Leftrightarrow\; L_{sym}(D^{1/2}x) = \lambda(D^{1/2}x).$$

The Laplacian spectrum can be associated with several graph invariants [15]. We note in particular the following result relating the number of connected components of an undirected graph G to the spectral properties of its associated graph Laplacian matrices L, L_{rw}, and L_{sym} [82].

Proposition 1. *Let G be an undirected graph with nonnegative weights. The multiplicity k of the eigenvalue 0 of L, L_{rw}, and L_{sym} is equal to the number of connected*

components $C_1, ..., C_k$ in the graph. For L and L_{rw}, the eigenspace of eigenvalue 0 is spanned by the indicator vectors $\mathbb{1}_{C_i}$ of these components. For L_{sym}, the eigenspace of eigenvalue 0 is spanned by the vectors $D^{1/2}\mathbb{1}_{C_i}$.

2.7 Connections with Related Techniques

With the above background, we now review Diffusion Maps and Locally Linear Embedding (LLE), two normalized output algorithms that can produce embeddings closely related to those of Laplacian Eigenmaps [4, 29].

2.7.1 Diffusion Maps

The diffusion maps algorithm provides a set of embeddings which aim to integrate and represent relationships between data points at different scales. Suppose we are given a data set $\Omega = \{x_1, ..., x_N\}$, and a kernel or weight matrix $W \in \mathbb{R}^{N \times N}$ that is symmetric ($W(x_i, x_j) = W(x_j, x_i)$) and positivity-preserving ($W(x_i, x_j) \geq 0$). The data elements could be points in a Euclidean space, representing, for example, a vector of measurements. Alternatively, they may be nodes of an explicitly specified graph, as in a constructed social or biological network. In all instances, the kernel can be specified to indicate any application-relevant similarity relationships, subject only to the above constraints. With the geometric information captured in the kernel, we can now define a Markov random walk on the data set viewed as a graph with data element nodes and kernel-weighted edges. In particular, the transition probability of going from node x_i to node x_j in a single step is

$$p_1(x_i, x_j) = \frac{W(x_i, x_j)}{\sum_{z \in \Omega} W(x_i, z)} = \frac{W(x_i, x_j)}{D(x_i, x_i)}, \tag{2.4}$$

where $D(x_i, x_i)$ represents the degree or connectivity strength of the node x_i, as specified in the appropriate entry of a diagonal matrix $D \in \mathbb{R}^{N \times N}$.

Let $P = D^{-1}W$ denote the $N \times N$ matrix whose (i, j) entry is the above-specified probability of transition from node x_i to node x_j. If we consider P^t, the matrix P raised to the power t, we record the corresponding probabilities, $p_t(x_i, x_j)$, of transition over t time steps. The Markov matrix P thus captures the first-order neighborhood structure of the data graph, while its iterates integrate higher order connectivity relationships by effectively running the random walk forward in time. The diffusion time t now becomes a convenient scale parameter in the analysis of the data structure. If we additionally assume that our data graph is connected, it can be shown that [18]

$$\lim_{t \to +\infty} p_t(x_i, x_j) = \frac{D(x_j, x_j)}{\sum_{z \in \Omega} D(z, z)} = \phi_0(x_j), \tag{2.5}$$

where ϕ_0 is the unique stationary distribution. Furthermore, the Markov chain is reversible, satisfying the detailed balance condition $\phi_0(x_i)p_1(x_i, x_j) = \phi_0(x_j)p_1(x_j, x_i)$. In this setting, it is natural to define the diffusion distance D_t between x_i and x_j as

$$D_t^2(x_i, x_j) = \|p_t(x_i, \cdot) - p_t(x_j, \cdot)\|_{1/\phi_0}^2 = \sum_{z \in \Omega} \frac{(p_t(x_i, z) - p_t(x_j, z))^2}{\phi_0(z)}. \tag{2.6}$$

Following the discussion in [18], we note some important attributes of the diffusion distance:

- The weighted L^2 distance specified above compares the transition probability distributions of nodes x_i and x_j, but with the weights $1/\phi_0(z)$ acting to specifically penalize deviations over regions of relatively low data density.

- The diffusion distance between two points will be small whenever they are highly connected over many paths in the graph. This will clearly be the case within densely connected regions, and the emphasis on aggregate connectivity tends to emphasize cluster structure in the data. In particular, at any given scale of analysis, clusters emerge as regions where the probability of 'escape' to less connected regions is low.

- The diffusion distance $D_t(x_i, x_j)$ effectively integrates information on all paths of length t connecting x_i and x_j. As such, it tends to be more robust with respect to noise perturbations, unlike the geodesic or shortest path distance. The latter distance can, for example, be spuriously compressed if noise-impacted data induces a non-intrinsic, 'short-circuit' connectivity between two points under consideration.

The Markov transition matrix described above has some specific spectral properties. In particular, the eigenvalues of P_t satisfy $1 = \lambda_0 \geq |\lambda_1| \geq ... \geq |\lambda_{N-1}|$. In addition, it can be shown that the diffusion distance can be expressed in terms of its eigenvalues and eigenvectors [18]:

$$D_t(x_i, x_j) = \left(\sum_{l \geq 1} \lambda_l^{2t} (\psi_l(x_i) - \psi_l(x_j))^2 \right)^{\frac{1}{2}}. \tag{2.7}$$

In the above, $\psi_l(x_i)$ and $\psi_l(x_j)$ denote eigenvector components corresponding to x_i and x_j, and we begin the summation from $l = 1$ since ψ_0 is constant (with all entries equal to one). Since the eigenvalues decay in magnitude, we can potentially approximate the diffusion distance D_t with a reduced number of summation terms. Specifically, given an accuracy $\delta > 0$, let $d(t)$ indicate the largest index l such that $|\lambda_l|^t > \delta|\lambda_1|^t$. The above summation over the first $d(t)$ nontrivial eigenvalues and eigenvectors will approximate the diffusion distance D_t up to a relative precision δ. In addition, if we define the following diffusion map

$$\Psi_t : x_i \mapsto \left(\lambda_1^t \psi_1(x_i), \lambda_2^t \psi_2(x_i), \ldots, \lambda_{d(t)}^t \psi_{d(t)}(x_i)\right)^T, \tag{2.8}$$

then we have the following result.

Proposition 2. *The diffusion map Ψ_t embeds the data into the Euclidean space $\mathbb{R}^{d(t)}$ so that in this space, the Euclidean distance is equal to the diffusion distance, up to relative accuracy δ, or equivalently,*

$$D_t^2(x_i, x_j) \simeq \sum_{l=1}^{d(t)} \lambda_l^{2t}(\psi_l(x_i) - \psi_l(x_j))^2 = \|\Psi_t(x_i) - \Psi_t(x_j)\|^2.$$

The extent of dimension reduction ultimately depends on both the scale parameter t and the decay of the eigenvalues, with the latter decay shaped by the connectivity structure of the kernel-prescribed data graph. Note that $P = D^{-1}W = I - L_{rw}$, so that the top eigenvectors of P used for the diffusion maps embedding correspond to the bottom eigenvectors of L_{rw}. As indicated in Section 2.6, the latter are also the bottom eigenvectors in the generalized eigenvalue problem $Lx = \lambda Dx$. The diffusion maps embedding is thus constructed from the same set of eigenvectors

used in Laplacian Eigenmaps, though with diffusion maps, these are now weighted with respect to the eigenvalues of the matrix P. If, for example, the data graph is connected and there are d clusters in the data, the d largest eigenvalues of P will be very close to 1, and the $(d-1)$-dimensional diffusion maps embedding will be quite similar to the $(d-1)$-dimensional Laplacian Eigenmaps embedding.

2.7.2 Locally Linear Embedding

The LLE algorithm begins with the intuition that, for a sufficiently smooth manifold, the local geometry is approximately linear within small neighborhoods. As a consequence, the mapping from the manifold to $\mathbb{R}^d$ is expected to be nearly linear within each such region. The algorithm proceeds in three steps to discover locally linear structure and represent the data so that this structure is approximately preserved in a low-dimensional embedding.

1. Identify Local Neighborhoods: for each x_i, find the k nearest neighbor set $N(i) = \{x_{i_1}, ..., x_{i_k}\}$.

2. Construct the Approximation Matrix W: Choose W_{ij} to minimize $\sum_{i=1}^{N} \|x_i - \sum_{j=1}^{k} W_{ij} x_{i_j}\|^2$, subject to the constraint $\sum_{j=1}^{k} W_{ij} = 1$.

   This is equivalent to orthogonally projecting each x_i onto the affine linear span of its neighborhood set x_{i_j}'s.

3. Compute the Embedding: Let $f_1, ..., f_d$ be the bottom d non-constant eigenvectors of the symmetric, positive semi-definite matrix $M = (I-W)^T(I-W)$. Embed in d-dimensional space using the map $x_i \to y_i = (f_1(i), f_2(i), ..., f_d(i))$.

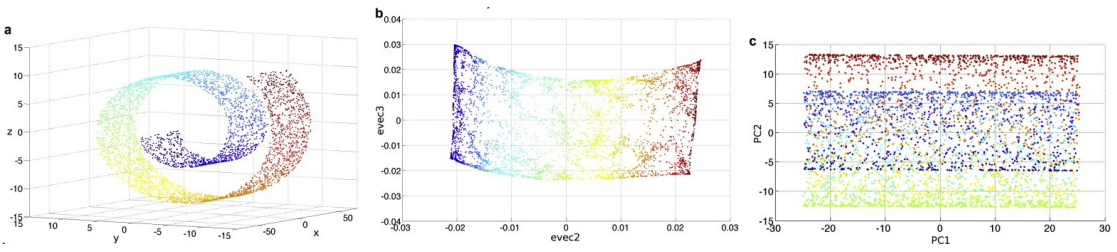


Figure 2.1: Nonlinear (Diffusion Maps) vs. linear (PCA) dimension reduction with the Swiss Roll data set.

The above embedding specifically minimizes the quantity

$$\sum_{i=1}^{N} \|y_i - \sum_{j=1}^{k} W_{ij} y_{i_j}\|^2,$$

to provide a d-dimensional point configuration with local geometry described by W [69]. If the neighborhoods used by LLE are in fact perfectly locally linear, it can be shown [4] that

$$(I - W)^T (I - W) f \approx \frac{1}{2} L^2 f.$$

Since the eigenvectors of $\frac{1}{2}L^2$ coincide with those of L, we note that the LLE embedding may be closely related to the LE embedding if, e.g., the above neighborhood assumptions are valid.

2.8 Examples

2.8.1 Linear vs. Nonlinear Dimension Reduction

We start with a simple and widely considered example to illustrate the distinctions between linear and nonlinear dimension reduction. In Figure 2.1, we show representations of the 'Swiss Roll' data set, which is derived by selecting 2000 points at

random from the corresponding flat, two-dimensional submanifold of $\mathbb{R}^2$. Although the Swiss Roll is essentially a rolled subset of the plane, PCA cannot possibly recover the underlying manifold structure, as it presumes a linear data organization. Projecting points to the best fitting plane, in this case, invariably forces together points that were far apart on the underlying data manifold, as can be seen from the co-mingled points of different colors. Diffusion maps, by contrast, can organize the points in a manner that broadly respects the manifold structure. Observe, however, that geodesic distances are not strictly preserved in the embedding. The the observed 'clumping' derives from the emphasis on local neighborhood preservation in Laplacian-based embeddings. The more detailed basis for this clustering effect will be considered in the following chapter.

2.8.2 An Illustrative Counterexample

Figure 2.2 is derived from [29], and aptly illustrates the fact that non-linear dimension reduction algorithms can be very sensitive to the input data, and may not always produce precisely faithful data representations. Plots (A) and (D) show identical initial sets of 3000 points, uniformly-sampled from the unit square, but scaled to the areas $[0, 81] \times [0, 41]$ and $[0, 81] \times [0, 39]$, respectively. Plots (B) and (E) show the outputs of Laplacian Eigenmaps for inputs (A) and (D), respectively, while plots (C) and (F) show the corresponding outputs for Diffusion Maps. For both algorithms, the parameters were kept the same for the two very similar inputs. Still, there is a pronounced difference in the output structure for the input data shown in

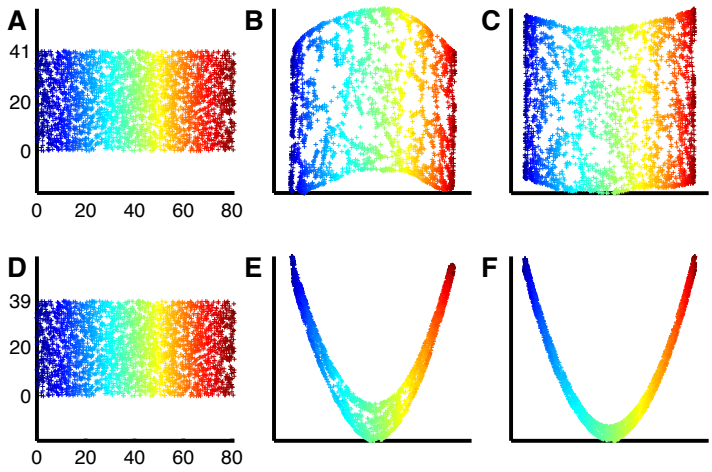


Figure 2.2: The above figure from [29] illustrates a result showing that Laplacian Eigenmaps (LE) and Diffusion Maps (DM) will produce an essentially one dimensional embedding of a two-dimensional point grid whenever its aspect ratio is greater than two. This is the case for the data in plot D, and the 'collapsed' embeddings produced by LE and DM are shown in plots E and F, respectively. These can be compared with the more faithful representations produced by the same algorithms in plots B and C, starting from the only slightly different data in plot A.

plot (D). While local neighborhood relationships are duly preserved, the simple, two-dimensional structure of the input data is not well-represented in either non-linear representation. The authors in fact show that these two normalized output methods will produce an essentially one dimensional embedding of a two-dimensional point grid whenever the aspect (width-to-height) ratio of the grid is greater than two. These examples show that the results of non-linear dimension reduction algorithms must be carefully assessed for preservation of structural features important to a particular application.

Chapter 3

Clustering Properties of Laplacian-Based Data Embeddings

The counterexample of Goldberg et al. [29] presented at the end of the preceding chapter shows that Laplacian-based embeddings do not always provide precisely faithful representations of manifold-sampled data. But in the context of data representations for learning, it is appropriate to consider if and how well other structural features are preserved under such embeddings. Perhaps the most fundamental concern is cluster structure preservation, as this impacts not just clustering, but also classification in supervised and semi-supervised formulations. In this chapter, we show that Laplacian-based embeddings do preserve cluster structure in reasonable settings. We begin with some motivating results, connecting Laplacian Eigenmaps to closely related spectral clustering techniques. We then proceed to a somewhat more illuminating perspective, which views 'real world data' with some cluster structure as a modest perturbation of idealized data with sharper group separation. From there, results based on matrix perturbation theory can be used to derive a precise statement about cluster structure preservation under Laplacian Eigenmaps. Some examples are finally presented to illustrate the theory.

3.1 Prior Results

Laplacian Eigenmaps can be related to spectral clustering techniques which approximate optimal graph partitionings [4, 82]. To recognize this, suppose we have a data graph with associated weight matrix W. We would like to partition our data into n dissimilar clusters of similar items. In graph theoretic terms, we seek to minimize the 'edge flow' between the clusters $C_1, ..., C_n$. If we let C_i^c denote the complement of C_i and set $W(C_i, C_i^c) = \sum_{u \in C_i, v \in C_i^c} W(u, v)$, an immediate idea is to try and minimize the total inter-cluster edge weight given by

$$cut(C_i, ..., C_n) = \frac{1}{2} \sum_{i=1}^{k} W(C_i, C_i^c).$$

This approach, while tractable, unfortunately often leads to unsatisfactory partitions that largely separate outliers. To capture the desirability of more balanced partitions, Shi and Malik [72] proposed the following normalized cut

$$Ncut(C_i, ...C_n) = \sum_{i=1}^{n} \frac{cut(C_i, C_i^c)}{vol(C_i)},$$

where $vol(C_i) = \sum_{u \in C_i} degree(u)$.

Minimizing $Ncut$ is in fact an NP-hard problem [72], but computations presented in [82] show that an approximation to its solution coincides with the solution to the eigenvalue problem considered in Laplacian Eigenmaps. To recognize this, let us represent the problem of minimizing $Ncut$ using the language of linear algebra. For the collection of n clusters formed from the elements of a set of N data points,

define the following cluster indicator vectors $h_j = (h_{1,j}...h_{N,j})^T$ by

$$h_{i,j} = \begin{cases} 1/\sqrt{vol(C_j)} & \text{if } v_i \in C_j, \\ 0 & \text{otherwise,} \end{cases}$$

for i = 1, ..., N; j = 1, ..., n. Let H be the matrix formed by considering the n indicator vectors h_j as columns. We can directly verify that $h_k^T h_l = 0$ for $k \neq l$, $h_j^T D h_j = 1$, and $h_j^T L h_j = cut(C_j, C_j^c)/vol(C_j)$. Minimizing $Ncut$ can thus be written as

$$\min_{C_i,...,C_n} \text{trace}(H^T L H) \quad \text{subject to} \quad H^T D H = I.$$

If we relax the requirement that the columns of H are discrete indicator vectors and substitute $U = D^{1/2}H$, we obtain

$$\min_{U \in \mathbb{R}^{N \times n}} \text{trace}(U^T D^{-1/2} L D^{-1/2} U) \quad \text{subject to} \quad U^T U = I.$$

This trace minimization problem is solved by the matrix U formed from the first n eigenvectors of $L_{sym} = D^{-1/2} L D^{-1/2}$ along its columns. Premultiplying by $D^{-1/2}$ the standard eigenvectors of L_{sym} gives the corresponding generalized eigenvectors of L, i.e., those eigenvectors that solve $Lu = \lambda Du$, which are used to construct the embedding in Laplacian Eigenmaps. From $D^{-1/2}U = H$, we have that these generalized eigenvectors also provide the solution to the relaxed formulation of the *Ncut* problem. In the normalized spectral clustering algorithm given by Shi and Malik [72], these eigenvectors are taken to be approximate cluster indicator vectors. K-means clustering of points formed from the rows of H is then used to separate the clusters.

The coordinates of the data points used for clustering coincide with those constructed for an $(n-1)$-dimensional Laplacian Eigenmaps embedding, except for the initial coordinate. This is because in LE, the initial coordinate would be excluded as it is derived from the Laplacian matrix eigenvector with eigenvalue zero, which would be constant for a connected data graph. Although these graph cut-based results motivate the clustering properties of Laplacian-based embeddings on a certain level, they still fail to provide precise guarantees of cluster structure preservation. This is because there is no guarantee relating the quality of the relaxed solution to that of the original *Ncut* problem [82].

3.2 Laplacian Eigenmaps with Perturbed Data

To better understand cluster structure preservation under Laplacian-based embeddings, we first examine how these embeddings are impacted by a perturbation. Let $\tilde{X} \in \mathbb{R}^{N\times m}$ denote a perturbed instance of our data set $X \in \mathbb{R}^{N\times m}$, i.e., $\tilde{X} = X + Z$, for $Z \in \mathbb{R}^{N\times m}$, with the entries of Z assumed to be bounded. As with X, we can construct a data adjacency matrix $\tilde{W} \in \mathbb{R}^{N\times N}$ from $\tilde{X}$, with entries given by

$$\tilde{W}_{i,j} = \begin{cases} \tilde{W}_{ij} = e^{\frac{-\|\tilde{x}_i - \tilde{x}_j\|^2}{\sigma}}, & \tilde{x}_i \text{ is one of the } k\text{-nearest neighbors of } \tilde{x}_j, \text{ or vice versa,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

From $\tilde{W}$, we can construct $\tilde{D}$ with $\tilde{D}_{ii} = \sum_{j=1}^{N} \tilde{W}_{ij}$, and ultimately $\tilde{L} = \tilde{D} - \tilde{W}$ and the closely related $\tilde{L}_{sym} = \tilde{D}^{-1/2}\tilde{L}\tilde{D}^{-1/2}$.

Note that the mapping from $\tilde{X}$ to $\tilde{W}$ is discontinuous, as a consequence of the neighborhood-based sparse adjacency matrix construction presented in (3.1). As a result, it is not possible to develop a bound based directly on the perturbation given by Z. Instead, we apply a result of Hunter and Strohmer [43] to characterize the effect of a Laplacian matrix perturbation on the Laplacian-based embedding. The overall approach builds on the work of Davis and Kahan in matrix perturbation theory [21]. In this setting, the notion of an *eigengap*, sometimes referred to as a *spectral gap*, appears frequently. For clarity, we note the following definition, and then proceed to a foundational result.

Definition 2. *Let $M \in \mathbb{R}^{N\times N}$ have the set of ordered eigenvalues $\lambda_1, ..., \lambda_N$, where the order could be, e.g., $\lambda_1 \leq .... \leq \lambda_N$ or $\lambda_1 \geq .... \geq \lambda_N$, as appropriate. We define the n^{th} eigengap to be $\gamma_n = |\lambda_n - \lambda_{n+1}|$.*

Theorem 3 (Hunter and Strohmer)**.** *Let $A = D^{-1/2}WD^{-1/2} \in \mathbb{R}^{N\times N}$ denote the normalized data adjacency matrix, and let $\tilde{A} = \tilde{D}^{-1/2}\tilde{W}\tilde{D}^{-1/2} \in \mathbb{R}^{N\times N}$ be the corresponding matrix constructed from perturbed data. For $1 \leq n \leq N$, let $V_n(i,\cdot)$, resp. $\tilde{V}_n(i,\cdot)$, be the i^{th} row of the matrix V_n, resp. $\tilde{V}_n$, with columns given by the n eigenvectors of A, resp. $\tilde{A}$, associated with the n largest eigenvalues. If $\gamma_n = \lambda_n - \lambda_{n+1} \geq \alpha$, and $\lambda_n \geq \alpha$, for $\alpha \in (0,1)$, then*

$$\|\tilde{V}_n(i,\cdot) - V_n(i,\cdot)Q\|_2 \leq (1+\sqrt{2})\frac{\sqrt{n}}{\alpha}\|A - \tilde{A}\|_F,$$

where Q is an orthogonal matrix that minimizes $\|\tilde{V}_n - V_nQ\|_F$.

The matrix perturbation $A - \tilde{A}$ in Theorem 3 is measured using the Frobenius

norm, which is defined for $M \in \mathbb{R}^{m \times n}$ as

$$\|M\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |M_{ij}|^2}.$$

Theorem 3 compares a data embedding based on the first n eigenvectors of A with one based on its perturbed version $\tilde{A}$, and shows that the corresponding embedded points are close, up to a unitary transform Q, if the perturbation is small and the n^{th} eigengap given by $\gamma_n = |\lambda_n - \lambda_{n+1}|$ is large. Applying Theorem 3, we can characterize the effect of a perturbation under the embedding constructed with Laplacian Eigenmaps. This embedding is obtained, as described in Section 2.4, using the eigenvectors associated with the n smallest *non-zero* eigenvalues in the generalized eigenvalue problem $Ly = \mu Dy$. To do so, we specifically establish a perturbation result for a more general Laplacian-based embedding, using n eigenvectors that solve $Ly = \mu Dy$, and are ordered with respect to non-decreasing values of μ. This embedding would include at least one such eigenvector with a corresponding eigenvalue of zero; see Section 2.6 for details.

Theorem 4. *Let $L, L_{sym}, D \in \mathbb{R}^{N \times N}$ and $\tilde{L}, \tilde{L}_{sym}, \tilde{D} \in \mathbb{R}^{N \times N}$ respectively denote the original and perturbed data-derived versions of the Laplacian, normalized Laplacian ($L_{sym} = D^{-1/2} L D^{-1/2}$), and diagonal degree matrices. For $1 \leq n \leq N$, let $U_n(i, \cdot)$ be the i^{th} row of the matrix $U_n \in \mathbb{R}^{N \times n}$, with columns given by n eigenvectors that solve $Ly = \mu Dy$, and are ordered by non-decreasing values of μ, starting from $\mu = 0$. Let $\tilde{U}_n(i, \cdot)$ be the i^{th} row of the matrix $\tilde{U}_n \in \mathbb{R}^{N \times n}$ with columns given by the corresponding set of eigenvectors solving $\tilde{L}y = \mu \tilde{D}y$. Finally, let $E(i) = \tilde{D}^{1/2}(i,i) - D^{1/2}(i,i)$ denote the point-specific connectivity perturbation. If $\gamma_n = \mu_{n+1} - \mu_n \geq \alpha$,*

and $\mu_n \leq (1-\alpha)$*, for* $\alpha \in (0,1)$*, then we have*

$$\|\tilde{U}_n(i,\cdot) - U_n(i,\cdot)Q\|_2 \leq D^{-1/2}(i,i)\left[(1+\sqrt{2})\frac{\sqrt{n}}{\alpha}\|\tilde{L}_{sym} - L_{sym}\|_F + \|E(i)\tilde{U}_n(i,\cdot)\|_2\right],$$

where Q *is an orthogonal matrix that minimizes* $\|\tilde{D}^{1/2}\tilde{U}_n - D^{1/2}U_nQ\|_F$*.*

Proof. Let $V_n \in \mathbb{R}^{N\times n}$ denote the matrix with columns given by the eigenvectors associated with the n largest eigenvalues of $A = D^{-1/2}WD^{-1/2}$, and let $V_n(i,\cdot)$ be the i^{th} row of V_n. Let $\tilde{V}_n$ and $\tilde{V}_n(i,\cdot)$ indicate the corresponding entities obtained starting from $\tilde{A} = \tilde{D}^{-1/2}\tilde{W}\tilde{D}^{-1/2}$. Our approach is to relate $\|\tilde{U}_n(i,\cdot) - U_n(i,\cdot)Q\|_2$ to the quantity that is bounded by Theorem 3, namely $\|V_n(i,\cdot)Q - \tilde{V}_n(i,\cdot)\|_2$.

Note that since $L_{sym} = D^{-1/2}LD^{-1/2} = I - A$, we have

$$L_{sym}y = \mu y \iff Ay = (1-\mu)y. \tag{3.2}$$

As a result, a set of eigenvectors associated with the n largest eigenvalues of A will also provide a set of eigenvectors corresponding to the n smallest eigenvalues of L_{sym}. We additionally have

$$Ly = \mu Dy \iff L_{sym}(D^{1/2}y) = \mu(D^{1/2}y). \tag{3.3}$$

Let $\mu_1 \leq ... \leq \mu_N$ denote the eigenvalues of L in the generalized eigenvalue problem $Ly = \mu Dy$, and let $\lambda_1 \geq ... \geq \lambda_N$ denote the eigenvalues of A. From (3.2) and (3.3), we have $\mu_i = 1 - \lambda_i$ and thus

$$\mu_{n+1} - \mu_n \geq \alpha, \;\; \mu_n \leq (1-\alpha) \iff \lambda_n - \lambda_{n+1} \geq \alpha, \;\; \lambda_n \geq \alpha$$

holds, as is necessary to apply Theorem 3.

From (3.2) and (3.3), and the construction of U_n, resp. $\tilde{U}_n$, with columns given by eigenvectors that solve $Ly = \mu Dy$, resp., $\tilde{L}y = \mu\tilde{D}y$, it follows that

$$V_n = D^{1/2}U_n, \;\; \tilde{V}_n = \tilde{D}^{1/2}\tilde{U}_n. \tag{3.4}$$

With $U_n(i,\cdot)$, resp. $\tilde{U}_n(i,\cdot)$, denoting the i^{th} row of the matrix U_n, resp. $\tilde{U}_n$, we thus have

$V_n(i,\cdot) = D^{1/2}(i,i)U_n(i,\cdot)$ and

$\tilde{V}_n(i,\cdot) = \tilde{D}^{1/2}(i,i)\tilde{U}_n(i,\cdot) = [D^{1/2}(i,i) + E(i)]\tilde{U}_n(i,\cdot)$.

We can accordingly write

$$\begin{aligned}
\|V_n(i,\cdot)Q - \tilde{V}_n(i,\cdot)\|_2 &= \|D^{1/2}(i,i)U_n(i,\cdot)Q - (D^{1/2}(i,i) + E(i))\tilde{U}_n(i,\cdot)\|_2 \\
&= \|D^{1/2}(i,i)U_n(i,\cdot)Q - D^{1/2}(i,i)\tilde{U}_n(i,\cdot) - E(i)\tilde{U}_n(i,\cdot)\|_2 \\
&\geq D^{1/2}(i,i)\|U_n(i,\cdot)Q - \tilde{U}_n(i,\cdot)\|_2 - \|E(i)\tilde{U}_n(i,\cdot)\|_2
\end{aligned}$$

or

$$\|U_n(i,\cdot)Q - \tilde{U}_n(i,\cdot)\|_2 \leq D^{-1/2}(i,i)[\|V_n(i,\cdot)Q - \tilde{V}_n(i,\cdot)\|_2 + \|E(i)\tilde{U}_n(i,\cdot)\|_2],$$

from which the result follows, upon applying the bound on $\|V_n(i,\cdot)Q - \tilde{V}_n(i,\cdot)\|_2$ presented in Theorem 3, and noting that $\|A - \tilde{A}\|_F = \|\tilde{L}_{sym} - L_{sym}\|_F$, and from (3.4), that

$\|\tilde{V}_n - V_nQ\|_F = \|\tilde{D}^{1/2}\tilde{U}_n - D^{1/2}U_nQ\|_F$. □

3.3 Cluster Structure Preservation with Laplacian Eigenmaps

Theorem 4 precisely characterizes the effect of a Laplacian matrix perturbation on the Laplacian-based embedding. This provides an avenue for understanding how

cluster structure is represented under such an embedding. Our approach is to regard realistic, e.g., noisy, data with cluster structure as a perturbation of data in which clusters are better separated and more apparent. For a basic example, consider the sets of four clusters in the plane shown in Figure 3.1. Between cluster set A and cluster set B, there is clearly a substantial difference in the inter-cluster distances. But, the data adjacency matrices W and $\tilde{W}$ constructed from cluster set A and cluster set B, respectively, will emphasize local neighborhood relationships. Thus, the corresponding Laplacian matrices L and $\tilde{L}$ will differ only modestly. In particular, $\tilde{L}$ will record associations that are not present in L, e.g., between neighboring points at the periphery of different clusters.

Viewing $\tilde{L}$ as a modest perturbation of L, we can apply Theorem 4 to understand the embedding of adjacent clusters in terms of the embedding of well-separated clusters. To do so, we first characterize the Laplacian-based embedding of well-separated clusters, e.g., clusters for which the data graph represented by W has cluster-specific connected components. Proposition 3 below shows that such separated clusters are mapped to orthogonally separated points, and are additionally organized according their intra-cluster connectivity. A related result has been established for a form of spectral clustering based on the eigenvectors of L_{sym} [60]. The n-dimensional embedding points associated with the latter method are additionally normalized to lie on the unit sphere in $\mathbb{R}^n$, yielding a different data representation from the Laplacian-based embedding analyzed in this section.

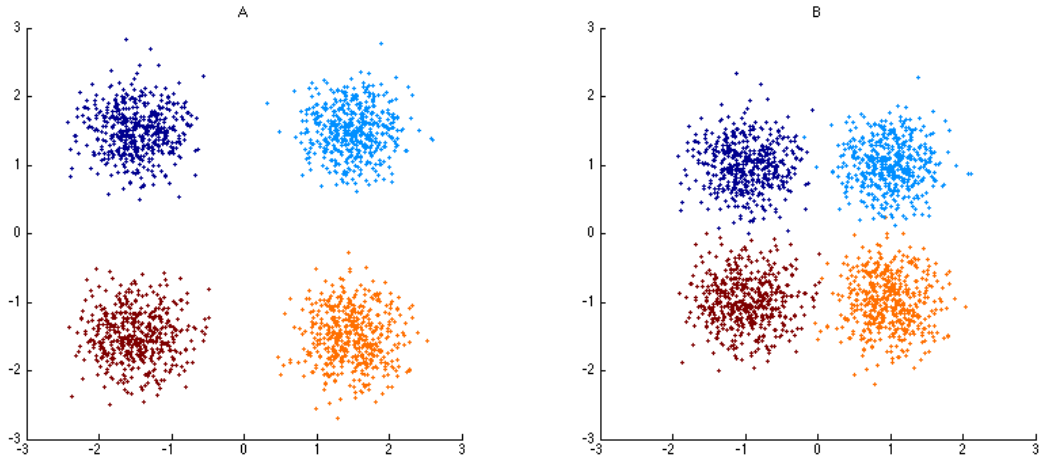


Figure 3.1: (A) Cluster Set A: four well-separated clusters. (B) Cluster Set B: four adjacent clusters, with points derived by translating points shown in Cluster Set A toward the origin.

Proposition 3. *Let $X \in \mathbb{R}^{N\times m}$ be a data set with elements $x_i = X(i,\cdot) \in \mathbb{R}^m$. Suppose that each point x_i is assigned to one of n clusters, $C_1, ..., C_n$, and furthermore, that the data adjacency matrix W is constructed so that*

$$\sum_{i\in C_k}\sum_{j\in C_l} W(i,j) = 0 \quad \textit{if} \quad k \neq l,$$

i.e., there are cluster-specific connected components in the underlying data graph, where we write $i \in C_k$ to indicate an index i such that $x_i \in C_k$. Let $L = D - W$ denote the Laplacian matrix constructed from W, with $D(i,i) = \sum_{j=1}^{N} W(i,j)$. Then L has exactly n eigenvectors $u_1, ..., u_n$ that solve $Lu_k = \mu Du_k$ with $\mu = 0$. If $U_n \in \mathbb{R}^{N\times n}$ is a matrix with u_k in the k^{th} column, and we use $U_n(i,\cdot)$ to denote the representation of x_i constructed from the i^{th} row of U_n, then there exist $c_1, ..., c_n \in \mathbb{R}^n$ such that

$$U_n(i,\cdot) = c_k \quad \textit{for all} \quad i \textit{ such that } x_i \in C_k,$$

i.e., we have cluster-specific embedding points, and furthermore,

$$c_k^T c_l = 0 \quad \textit{if} \quad k \neq l.$$

In addition, if the u_k are scaled so that $u_k^T D u_k = 1$, then

$$\|c_k\|_2 = \frac{1}{\sqrt{Vol(C_k)}},$$

where $Vol(C_k) = \sum_{i \in C_k} D(i,i)$.

Proof. From the construction of W such that the data graph has cluster-specific connected components, it follows from the basic properties of Laplacian matrices that L has exactly n eigenvectors $u_1, ..., u_n$ that solve $Lu_k = \mu D u_k$ with $\mu = 0$; see Proposition 1 in Section 2.6 for details.

Since we have, for $L_{sym} = D^{-1/2} L D^{-1/2}$,

$$L(D^{-1/2} v_k) = \mu D(D^{-1/2} v_k) \iff L_{sym} v_k = \mu v_k, \tag{3.5}$$

we can obtain the eigenvectors u_k that solve $Lu_k = \mu D u_k$ with $\mu = 0$ by pre-multiplying by $D^{-1/2}$ the standard eigenvectors v_k that solve $L_{sym} v_k = \mu v_k$ with $\mu = 0$.

To proceed further, note that we can assume, without loss of generality, that the points $x_1, ...x_N$ are ordered with respect to cluster membership. Accordingly, the matrix L_{sym} will be block diagonal with cluster-specific component Laplacians

$$L_{sym} = \begin{bmatrix} L_{sym}^{(1)} & & & \\ & L_{sym}^{(2)} & & \\ & & \ddots & \\ & & & L_{sym}^{(n)} \end{bmatrix}.$$

One can verify that the eigenspace of eigenvalue zero for L_{sym} is spanned by n eigenvectors given by the columns of

$$D^{1/2}U = D^{1/2}\begin{bmatrix} 1^{(1)} & & & \\ & 1^{(2)} & & \\ & & \ddots & \\ & & & 1^{(n)} \end{bmatrix} \in \mathbb{R}^{N\times n},$$

where each $1^{(k)}$ denotes an all ones column vector with length equal to $|C_k|$. These eigenvectors are typically normalized, and we can represent a set of eigenvectors spanning the zero eigenspace of L_{sym} as the columns of

$$D^{1/2}USR \in \mathbb{R}^{N\times n},$$

where $S \in \mathbb{R}^{n\times n}$ is a diagonal scaling matrix given by

$$S = \begin{bmatrix} \frac{1}{\sqrt{Vol(C_1)}} & & & \\ & \frac{1}{\sqrt{Vol(C_2)}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{Vol(C_n)}} \end{bmatrix},$$

and $R \in \mathbb{R}^{n\times n}$ is any orthogonal matrix.

It now follows from (3.5) that $U_n = USR \in \mathbb{R}^{N\times n}$ is a matrix whose columns contain the eigenvectors u_k that solve $Lu_k = \mu Du_k$ with $\mu = 0$. From the construction of U_n, it is clear that each embedded point $U_n(i,\cdot)$ is given by a cluster-specific row of $SR \in \mathbb{R}^{n\times n}$. Let $c_k \in \mathbb{R}^n$ denote the k^{th} row of SR. Since the rows of SR are orthogonal, we have

$$c_k^T c_l = 0 \quad \text{if} \quad k \neq l.$$

Finally, from the construction of S and $R^T R = I$, it follows that

$$\|c_k\|_2 = \frac{1}{\sqrt{Vol(C_k)}}.$$

□

Proposition 3 shows that for a data set X with well-separated clusters $C_1, ...C_n$, the cluster structure is accentuated in the embedded space, with $x_i \in C_k$ collapsing to a single cluster-specific point $c_k \in \mathbb{R}^n$ in the embedding. Combining Theorem 4 and Proposition 3 we obtain the following result.

Proposition 4. *Let $X \in \mathbb{R}^{N\times m}$ be a data set with elements $x_i = X(i,\cdot) \in \mathbb{R}^m$. Suppose that each point x_i is assigned to one of n clusters, $C_1, ..., C_n$, and furthermore, that the data adjacency matrix $W \in \mathbb{R}^{N\times N}$ is constructed so that there are cluster-specific connected components in the underlying data graph. Let $\tilde{X} \in \mathbb{R}^{N\times m}$ be a perturbation of X, with $\tilde{x_i} = \tilde{X}(i,\cdot) \in \mathbb{R}^m$ retaining the cluster assignment of x_i, and let $\tilde{W} \in \mathbb{R}^{N\times N}$ denote the data adjacency matrix constructed from $\tilde{X}$. Let $L, L_{sym}, D \in \mathbb{R}^{N\times N}$ and $\tilde{L}, \tilde{L}_{sym}, \tilde{D} \in \mathbb{R}^{N\times N}$ respectively denote the Laplacian, normalized Laplacian ($L_{sym} = D^{-1/2}LD^{-1/2}$), and diagonal degree matrices constructed from W and $\tilde{W}$. Let $\tilde{U}_n(i,\cdot)$ be the i^{th} row of the matrix $\tilde{U}_n \in \mathbb{R}^{N\times n}$ with columns given by n eigenvectors that solve $\tilde{L}y = \mu\tilde{D}y$ and are ordered by non-decreasing values of μ, starting from $\mu = 0$. Let $U_n(i,\cdot)$ be the i^{th} row of the matrix $U_n \in \mathbb{R}^{N\times n}$, with columns given by the corresponding set of eigenvectors solving $Ly = \mu Dy$. Let $\tilde{y}_i = \tilde{U}_n(i,\cdot)$ denote the point representing $\tilde{x}_i$ in the n-dimensional Laplacian-based embedding constructed as described from $\tilde{L}$, and let $c_k \in \mathbb{R}^n$ indicate the cluster-specific embedding point for $x_i \in C_k$. Finally, let $E(i) = \tilde{D}^{1/2}(i,i) - D^{1/2}(i,i)$*

denote the connectivity perturbation associated with $\tilde{x}_i$. *If* $\gamma_n = \mu_{n+1} - \mu_n \geq \alpha$, *and* $\mu_n \leq (1-\alpha)$, *for* $\alpha \in (0,1)$, *then we have*

$$\|\tilde{y}_i - c_k Q\|_2 \leq D^{-1/2}(i,i)\left[(1+\sqrt{2})\frac{\sqrt{n}}{\alpha}\|\tilde{L}_{sym} - L_{sym}\|_F + \|E(i)\tilde{U}_n(i,\cdot)\|_2\right],$$

where Q *is an orthogonal matrix that minimizes* $\|\tilde{D}^{1/2}\tilde{U}_n - D^{1/2}U_n Q\|_F$.

Proof. Note that in the theorem statement, we set $\tilde{y}_i = \tilde{U}_n(i,\cdot)$. From Proposition 3, we have $U_n(i,\cdot) = c_k$ for $x_i \in C_k$, whenever the data adjacency matrix W associated with X has cluster-specific connected components. With these substitutions, the result follows directly from Theorem 4. □

Basic results from spectral graph theory show that the eigengap γ_n is large whenever n coherent clusters are present, i.e., no 'bottleneck edges' exist that can be cut to subdivide existing clusters in a balanced way [15]. Proposition 4 thus shows that embedded points derived from the less separated clusters in $\tilde{X}$ deviate from the cluster-specific points associated with the well-separated clusters in X according to a bound governed by the initial cluster coherence and point-specific connectivity strength in X, and notably, the inter-cluster connectivity that emerges in $\tilde{X}$. These attributes are indicated by α, the magnitude of the lower bound on the eigengap γ_n, and for a point $\tilde{x}_i$, the factor $D^{-1/2}(i,i)$, and the terms $\|\tilde{L}_{sym} - L_{sym}\|_F$ and $E(i)$.

These ideas are illustrated by the example shown in Figure 3.2. We start with three clusters concentrated on roughly semi-circular curves in the plane. Although the clusters are well-separated, their boundaries are non-convex and k-means clustering will not reveal the visually apparent cluster structure. However, from the

preceding discussion, it follows that a reasonable construction of the Laplacian matrix will embed the points in three-dimensional space, with cluster-associated points mapped to mutually orthogonal points. With Laplacian Eigenmaps, the first, constant eigenvector associated with eigenvalue zero is typically discarded. The embedding in terms of the second and third eigenvectors is thus a projection of the orthogonally separated, cluster-mapped points onto the plane. These points are still well-separated, and k-means clustering in the embedded space has no difficulty exposing the true cluster structure.

Figure 3.3, shows a more realistic case of less clearly separated clusters. In the presented theoretical framework, the data-derived Laplacian matrix is regarded as a perturbed version of the Laplacian for the previous case of well-separated clusters. In keeping with the results relating the original and perturbed case embeddings, note that the cluster structure is still clearly revealed in the embedded space, with only modest distortion relative to the original, well-separated case. K-means, or any other reasonable clustering algorithm, can still readily discern the clusters. These simple examples still illustrate a notable advantage of clustering in the Laplacian-mapped space. Namely, the ability to detect a broader range of cluster structure, including non-convex patterns.

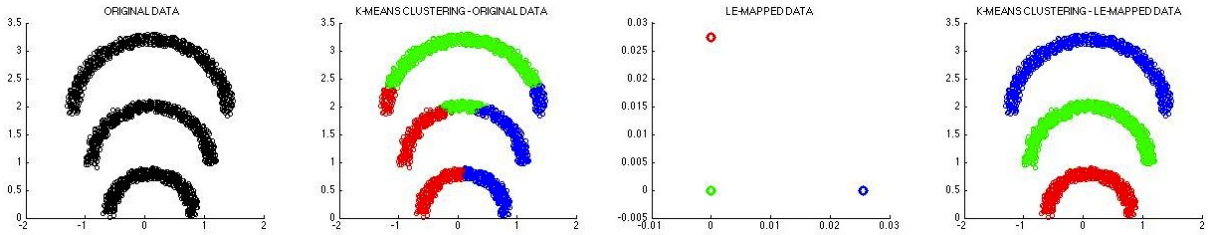


Figure 3.2: Clustering of well-separated non-convex clusters in original and Laplacian-mapped spaces.

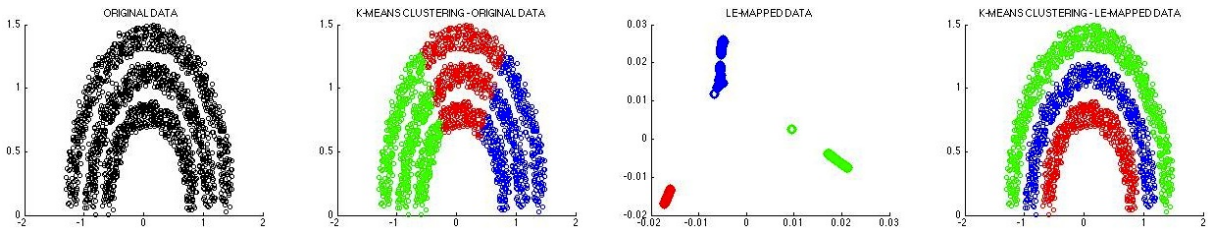


Figure 3.3: Clustering of less separated non-convex clusters in original and Laplacian-mapped spaces.

Chapter 4

Information Fusion

4.1 Introduction

In this chapter, we present techniques for two distinct types of data fusion. In the first instance, we have a collection of data sets $X_1, ..., X_M$, with each data set providing a different 'view' of a fixed set of elements. Points in each data set can be related with data set-specific kernels $W_1, ..., W_M$, and a direct approach for information fusion is to combine the latter kernels into a 'fusion kernel' W_F. The fusion kernel W_F can then be applied within the Laplacian-based data representation framework introduced in Chapter 2, and further characterized in Chapter 3.

This sort of *multi-kernel information fusion* is directly applicable when we aim to combine data sets providing different observations of a single collection of entities. In this context, there is a clear, bijective correspondence between elements of the various data sets $X_1, ..., X_M$. While this fusion model fits numerous applications in which multiple data sets provide alternative measurements of the same entities, it is not appropriate for instances in which the aim is to relate data describing fundamentally distinct sets of elements. Suppose instead that we have two data sets X and Y, and there is no fully specified bijective correspondence between their elements. At the same time, there are known or expected relationships between at least certain elements across the data sets. At the level of the underlying data graphs

Γ_X and Γ_Y, we might expect subgraphs with similar structure and corresponding elements. In Section 4.3, we present an algorithm for organizing the related, but heterogeneous elements underlying data sets X and Y in a single joint embedding. Our starting point is an extension of the Diffusion Maps framework developed by Coifman and Hirn [17], which allows us to identify points that can be used to 'align' the data graphs Γ_X, Γ_Y and construct a joint embedding.

4.2 Multi-Kernel Information Fusion

Suppose we have a collection of data sets $X_1, ..., X_M$, with $X_l \in \mathbb{R}^{N \times D_l}$ for $l = 1, ..., M$. Each data set provides a different collection of measurements for a fixed set of N elements, with the i^{th} element in the data set X_l denoted by $X_l(i, \cdot) \in \mathbb{R}^{D_l}$. With data set-specific choices for the kernel bandwidth and neighborhood set size parameters, indicated by σ_l and k_l respectively, we can construct data adjacency matrices $W_l \in \mathbb{R}^{N \times N}$, with entries given by

$$W_l(i, j) = e^{\frac{-\|X_l(i,\cdot) - X_l(j,\cdot)\|^2}{\sigma_l}},$$

if $X_l(i, \cdot)$ is among the k_l nearest neighbors of $X_l(j, \cdot)$, or vice versa, and $W_l(i, j) = 0$ otherwise. More generally, we can suppose that W_l is a symmetric, non-negative-valued kernel matrix with entries relating pairs of elements in X_l.

Define

$$\mathcal{W} = \{W \in \mathbb{R}^{N \times N} \mid W^T = W \text{ and } W(i, j) \geq 0 \text{ for all } i, j \in 1, ..., N\}$$

to be the set of $N \times N$ kernel matrices. A natural approach for information fusion is to apply a function $F : \mathcal{W}^M \to \mathcal{W}$ to combine the set of kernels given by $W_1, ..., W_M$

into a single kernel W_F. This 'fusion kernel' W_F can then be applied within the Laplacian-based data representation framework, with the significant eigenvectors of the fusion Laplacian matrix L_F used to construct an embedding that integrates features provided by the data sets $X_1, ...X_M$. The precise specification of F, i.e., the construction of the fusion kernel W_F, is to a substantial extent application dependent. In this section, we briefly outline some broad approaches, together with a few associated observations. A particular application of this multi-kernel information fusion approach is presented in Section 5.2.

Two direct approaches for kernel integration are to construct linear combinations of data set-specific kernels, i.e.,

$$W_{F_s} = \sum_{l=1}^{M} \alpha_l W_l,$$

or weighted pointwise products of kernels, i.e., form W_{F_p}, with entries given by

$$W_{F_p}(i,j) = \prod_{l=1}^{M} \alpha_l W_l(i,j).$$

An 'unbiased' fusion approach would set $\alpha_l = 1$ for $l = 1, ..., M$ in either case. The two approaches induce different types of data graphs, and thus different sorts of integrative embeddings. In particular, the pointwise product kernel W_{F_p} will emphasize data element relationships that are strong across the full collection of data sets. Weaker associations recorded by two or more kernels will negatively reinforce one another to produce further reduced entries, and clearly if $W_l(i,j) \approx 0$ in even a single kernel, we will have $W_{F_p}(i,j) \approx 0$. Where there is some heterogeneity across the data sets X_l, use of W_{F_p} often yields data graphs with multiple connected components, grouping data elements that are consistently related across the collection

of data sets. In some application contexts, these may directly represent meaningful clusters, but often, they also group large numbers of data elements which must be further sub-divided using connected component-specific Laplacian embeddings. If the specific application does not suggest 'disconnecting' points that are not substantially related across the full collection of data sets $X_1, ...X_M$, it is perhaps more appropriate to apply the linear combination-based fusion kernel W_{F_s}.

With both the linear combination and the pointwise product-based kernel fusion approaches, an additional detail is selection of the weights α_l. As noted, a direct approach is to set these equally to one, or otherwise, to weight particular kernels based on application-specific considerations, such as the relative importance or reliability of their associated data sets. Alternatively, for the particular case of fusion with two kernels, W_1 and W_2, a direct grid search could be used to select a pair of weights that maximizes a suitable measure of embedding quality. In particular, if $Y_{\alpha_1,\alpha_2} \in \mathbb{R}^{N\times d}$ denotes an embedding derived from the fusion kernel $\alpha_1 W_1 + \alpha_2 W_2$, and $Q : \mathbb{R}^{N\times d} \to \mathbb{R}$ is a function specifying the embedding quality, we would select

$$\underset{(\alpha_1,\alpha_2)\in A_1\times A_2}{\arg\max} Q(Y_{\alpha_1,\alpha_2}),$$

where A_1 and A_2 are sets of candidate values for α_1 and α_2, respectively. The same approach could clearly be applied to optimize the weights in a pointwise product of two kernels W_1 and W_2. Note that the function Q in the described scheme could generally represent a computational procedure that returns a measure of embedding quality. For example, if the aim is to select a kernel fusion-based embedding that maximizes cluster structure, Q could apply a clustering algorithm to the embedded

data and return a measure of the cluster quality.

4.3 Joint Embeddings for Heterogeneous Data Fusion

In this section, we present an algorithm for a more general sort of data fusion, where the aim is to co-organize two data sets, X and Y, recording measurements of distinct but at least partially related underlying elements. In particular, the algorithm constructs a joint embedding which further reveals these relationships between the data sets, while attempting to preserve the intrinsic relations between elements within each data set. Our starting point is an extension of the diffusion maps framework developed by Coifman and Hirn [17]. The authors consider a *fixed* set of data elements, and a collection of data sets X_α parameterized by $\alpha \in \mathcal{I}$. While the data set elements are unchanged over the X_α, the relationships between these elements can vary. These changes are captured in the data, and modeled as alterations to the derived data graphs Γ_α. For a fixed α, a diffusion map can be constructed as described in [18], to embed the data according to the geometry represented by the data graph Γ_α; see also Subsection 2.7.1. But, for different parameters α and β, the embeddings of X_α and X_β are to distinct spaces, and direct comparisons between mapped data set elements across the parameterized contexts are not possible. Coifman and Hirn generalize the diffusion maps framework to define formulas for the distance between points in different embeddings. They additionally define a mapping from one embedding to another. In this joint embedding, diffusion distance relationships within each data set are preserved, while corresponding points derived

from different data sets are organized according to a generalized diffusion distance which measures how much the local subgraph around each point changes over the parameter space.

The diffusion maps framework for changing data developed by Coifman and Hirn presumes an explicit bijection between data set elements across the X_α. We adapt this approach to the fusion of heterogeneous data sets X and Y by starting with an initial, tentative bijection between related elements across the data sets. This bijection can be constructed using prior knowledge or some objective analysis of the data. A Coifman-Hirn algorithm-based joint embedding of the data, constructed with respect to this bijection, is then used to identify corresponding points with genuinely similar local neighborhood structure. If these paired points are sufficiently representative of the two data sets, they can be used as 'hooks' to meaningfully align their respective data graphs and construct a joint embedding. We consider the paired points to be sufficiently representative of their associated data sets if each provides a frame, i.e., at least a spanning set, for its larger data set. Our approach starts with the Coifman-Hirn embedding of the paired data. Other data points are then embedded as linear combinations of the embedded paired points, with specific linear combination weights derived from the data set-specific frame reconstruction coefficients.

To present this approach for jointly embedding heterogeneous data in more detail, we start by reviewing essential elements of the Coifman-Hirn generalization of diffusion maps to support changing data. After a summary of frames and techniques for constructing sparse, frame-based data representations, we describe the

joint embedding algorithm. Details of the algorithm, such as the construction of the initial bijection relating elements of the data sets, are application dependent and amenable to some adjustment. A particular case study is presented in Subsection 5.3.4.

4.3.1 Diffusion Maps for Changing Data

The Coifman-Hirn generalization of diffusion maps assumes that the data points are drawn from a single measure space (X, μ), which changes over the parameter space $\mathcal{I}$. There are no restrictions on $\mathcal{I}$, i.e., it can be discrete, continuous, or arbitrary, but the underlying point distribution represented by the measure μ is assumed to be fixed. The change in (X, μ) being considered is instead with respect to the relationships between points. This evolution is precisely described by a family of metrics $d_\alpha : X \times X \to \mathbb{R}$, and for each $\alpha \in \mathcal{I}$, there is thus a metric measure space $X_\alpha = (X, \mu, d_\alpha)$, for which one can construct a kernel $k_\alpha : X \times X \to \mathbb{R}$, together with a weighted data graph $\Gamma_\alpha \triangleq (X, k_\alpha)$. For the results summarized below, it is further assumed that the kernel k_α is positive definite and symmetric for all $\alpha \in \mathcal{I}$.

To allow comparison of geometric structures of X across the parameter space $\mathcal{I}$, the diffusion maps framework reviewed in Subsection 2.7.1 is adjusted. In particular, for each parameter $\alpha \in \mathcal{I}$, the density $m_\alpha : X \to \mathbb{R}$ is defined as

$$m_\alpha(x) \triangleq \int_X k_\alpha(x, y)\, d\mu(y), \quad \text{for all } \alpha \in \mathcal{I}, x \in X,$$

with the assumption that $m_\alpha \in L^1(X, \mu)$, for all $\alpha \in \mathcal{I}$, and $m_\alpha(x) > 0$ for all $\alpha \in \mathcal{I}$

and $x \in X$. From here, the parameterized kernel $a_\alpha : X \times X \to \mathbb{R}$ is given by

$$a_\alpha(x,y) \triangleq \frac{k_\alpha(x,y)}{\sqrt{m_\alpha(x)}\sqrt{m_\alpha(y)}}, \text{ for all } \alpha \in \mathcal{I}, (x,y) \in X \times X,$$

together with its corresponding integral operator $A_\alpha : L^2(X,\mu) \to L^2(X,\mu)$, where

$$(A_\alpha f)(x) \triangleq \int_X a_\alpha(x,y) f(y) d\mu(y), \text{ for all } \alpha \in \mathcal{I}, f \in L^2(X,\mu). \tag{4.1}$$

Finally, let $a_\alpha^{(t)}$ denote the kernel of the integral operator $A_\alpha^{(t)}$, which allows a multiscale analysis of the data based on advancing the Markov chain by t time steps.

With these definitions, Coifman and Hirn introduce the following generalized diffusion distance.

Definition 3 (Dynamic Diffusion Distance)**.** *Let $x_\alpha \triangleq (x,\alpha) \in X \times \mathcal{I}$. For each diffusion time $t \in \mathbb{N}$, the dynamic diffusion distance $D^{(t)} : (X \times \mathcal{I}) \times (X \times \mathcal{I}) \to \mathbb{R}$ is defined as*

$$\begin{aligned} D^{(t)}(x_\alpha, y_\beta)^2 &\triangleq \|a_\alpha^{(t)}(x,\cdot) - a_\beta^{(t)}(y,\cdot)\|^2_{L^2(X,\mu)} \\ &= \int_X (a_\alpha^{(t)}(x,u) - a_\beta^{(t)}(y,u))^2 \, d\mu(u). \end{aligned}$$

We can understand the diffusion distance by fixing a point $x \in X$, and considering the function $a_\alpha^{(t)}(x,\cdot)$ together with the data graph Γ_α. If a unit of mass is placed on the node x and allowed to diffuse over Γ_α, the quantity of mass that has spread from x to y over t time steps is proportional to $a_\alpha^{(t)}(x,y)$. The time t diffusion distance between x and y thus compares the pattern of diffusion centered at x with the corresponding pattern centered at y. These patterns are based on the local connectivity structure around these points, so if the neighborhood of x_α is similar to

that of y_β, their diffusion distance will be small. In particular, the dynamic diffusion distance between x_α and x_β precisely measures the change in neighborhood structure around the point x between the two parameterized contexts.

Under some mild assumptions indicated in [17], it can be shown that the operators A_α are compact, positive definite, and self-adjoint. From the Spectral Theorem, it thus follows that each operator A_α has a countable set of positive eigenvalues and orthonormal eigenfunctions, and that the latter eigenfunctions provide a basis for $L^2(X, \mu)$. Let $\{\lambda_\alpha^{(i)}\}_{i\geq 1}$ and $\{\psi_\alpha^{(i)}\}_{i\geq 1}$ denote these eigenvalues and a set of orthonormal eigenfunctions, respectively. Coifman and Hirn show that as with the original, single data set diffusion distance, the generalized diffusion distance can be expressed in terms of the spectral decompositions of the relevant operators. In particular:

$$D^{(t)}(x_\alpha, y_\beta)^2 = \sum_{i\geq 1}(\lambda_\alpha^{(i)})^{2t}\psi_\alpha^{(i)}(x)^2 + \sum_{j\geq 1}(\lambda_\beta^{(j)})^{2t}\psi_\beta^{(j)}(y)^2$$
$$-2\sum_{i,j\geq 1}(\lambda_\alpha^{(i)})^t(\lambda_\beta^{(j)})^t\psi_\alpha^{(i)}(x)\psi_\beta^{(j)}(y)\langle\psi_\alpha^{(i)}, \psi_\beta^{(j)}\rangle_{L^2(X,\mu)}.$$

Note, as shown in [17], that we have

$$1 = \lambda_\alpha^{(1)} \geq \lambda_\alpha^{(2)} \geq \lambda_\alpha^{(3)} \geq ...,$$

with $\lambda_\alpha^{(i)} \to 0$ as $i \to \infty$. It thus follows that the diffusion distance can be well approximated by a small number of eigenvalues and eigenfunctions of the operators A_α and A_β, if their spectra decay sufficiently fast.

For the parameter α and diffusion time t, the diffusion map $\Psi_\alpha^{(t)} : X \to \ell^2$ is defined as

$$\Psi_\alpha^{(t)}(x) \triangleq \left((\lambda_\alpha^{(i)})^t\psi_\alpha^{(i)}(x)\right)_{i\geq 1}. \tag{4.2}$$

While each diffusion map $\Psi_\alpha^{(t)}$ sends X into an ℓ^2 space specific to α, the generalized diffusion distance allows distances to be meaningfully computed between two diffusion embeddings. Conveniently, it is possible to map one such embedding into another using an operator similar to the change of basis operator. In particular, define $O_{\beta\to\alpha} : \ell^2 \to \ell^2$ as

$$O_{\beta\to\alpha} v \triangleq \left(\sum_{j\geq 1} v[j] \langle \psi_\alpha^{(i)}, \psi_\beta^{(j)} \rangle_{L^2(X,\mu)} \right)_{i\geq 1}, \text{ for all } v \in \ell^2. \tag{4.3}$$

Using the operator $O_{\beta\to\alpha}$, one can construct a joint embedding which relates points originally in separate embeddings according to the generalized diffusion distance, while also preserving the initial, intra-embedding diffusion distances.

4.3.2 Frames and Sparse Data Representation

Frames, introduced by Duffin and Schaeffer in 1952 [24], generalize bases to provide robust and flexible representations of vectors. A basis for a finite dimensional Hilbert space $\mathbb{H}$ is a set that can be used to *uniquely* represent each element of $\mathbb{H}$. A frame for $\mathbb{H}$, on the other hand, may include elements which are linearly dependent. Such overcomplete sets allow for an infinite number of representations of a given element in $\mathbb{H}$. In signal coding applications, this underlying redundancy often enables resilience in the face of erasures and other sorts of errors [12, 31]. More generally, as overcomplete systems, frames can be flexibly designed to concentrate representation elements in specific regions of the data space, enhancing resolution of application-relevant data features. This aspect of frames is useful in the context of the joint embedding algorithm presented in Subsection 4.3.3 .

To introduce frames in general terms, let $\mathcal{I}$ be a possibly infinite, but countable, index set. A sequence $\mathcal{F} = \{f_i\}_{i\in\mathcal{I}}$ in a separable Hilbert space $\mathbb{H}$ is a *frame* for $\mathbb{H}$ if there exist $0 < A \leq B < \infty$ such that

$$A\|f\|^2 \leq \sum_{i\in\mathcal{I}} |\langle f, f_i\rangle|^2 \leq B\|f\|^2 \qquad \forall f \in \mathbb{H}. \tag{4.4}$$

A and B are *lower and upper frame bounds*, and when these can be chosen as $A = B$, $\mathcal{F}$ is called an A-*tight frame*, with the particular case of $A = B = 1$ commonly referred to as a *Parseval* frame.

There are three fundamental operators in frame theory:

- The *analysis operator* $T_\mathcal{F} : \mathbb{H} \mapsto \ell_2(\mathcal{I})$ given by $T_\mathcal{F} f = \{\langle f, f_i\rangle\}_{i\in\mathcal{I}}$ maps a signal $f \in \mathbb{H}$ to the *representation space* $\ell_2(\mathcal{I})$.

- A mapping from the representation space back to $\mathbb{H}$ is provided by the *synthesis operator*, which is defined as the adjoint operator $T^*_\mathcal{F} : \ell_2(\mathcal{I}) \mapsto \mathbb{H}$ given by $T^*_\mathcal{F}(\{c_i\}_{i\in\mathcal{I}}) = \sum_{i\in\mathcal{I}} c_i f_i$.

- Composing $T_\mathcal{F}$ and $T^*_\mathcal{F}$ gives the *frame operator* $S_\mathcal{F} : \mathbb{H} \mapsto \mathbb{H}$, $S_\mathcal{F} f = T^*_\mathcal{F} T_\mathcal{F} f = \sum_{i\in\mathcal{I}} \langle f, f_i\rangle f_i$, which is positive-definite, self-adjoint, and invertible.

Given a frame $\mathcal{F} = \{f_i\}_{i\in\mathcal{I}}$, we have the reconstruction formula

$$f = \sum_{i\in\mathcal{I}} \langle f, f_i\rangle S^{-1}_\mathcal{F} f_i = \sum_{i\in\mathcal{I}} \langle f, S^{-1}_\mathcal{F} f_i\rangle f_i \qquad \forall f \in \mathbb{H}, \tag{4.5}$$

where $\{{\mathcal{S}_\mathcal{F}}^{-1} f_i\}_{i\in\mathcal{I}}$ is the *canonical dual frame*. When $\mathcal{F}$ is redundant, there exist infinitely many dual frames $\{\tilde{f}_i\}_{i\in\mathcal{I}}$, but the canonical dual will satisfy the least

squares property among all dual frames $\{\tilde{f}_i\}_{i\in\mathcal{I}}$, i.e.,

$$\sum_{i\in\mathcal{I}} |\langle f, S_{\mathcal{F}}^{-1} f_i\rangle|^2 \leq \sum_{i\in\mathcal{I}} |\langle f, \tilde{f}_i\rangle|^2 \qquad \forall f \in \mathbb{H}. \tag{4.6}$$

Note that for A-tight frames, we have $S^{-1} = \frac{1}{A}\mathbb{I}$, giving the simple reconstruction formula $f = \frac{1}{A} T_{\mathcal{F}}^* T_{\mathcal{F}} f$, for all $f \in \mathbb{H}$. In this case, we see that a frame-based decomposition is essentially as efficient and convenient as one using an orthonormal basis. If $\mathbb{H}$ is an n-dimensional Hilbert space, a frame is easy to characterize: the k-element collection of vectors $\{f_i\}_{i=1}^k$ is a frame for $\mathbb{H}$ if and only if it spans $\mathbb{H}$. The *redundancy* of such a frame is the quantity $\frac{k}{n}$. More information on the robustness and flexibility afforded by frames can be found in [47, 48]. For derivation of the basic properties summarized above, see also [14].

While the reconstruction formula presented in Equation 4.5 yields a set of representation coefficients with minimal ℓ^2 norm, this may not be the most desirable representation in many application contexts. Suppose that $\{f_i\}_{i=1}^k$ is a frame for $\mathbb{F}^n$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Given $f \in \mathbb{F}^n$, we may wish to reconstruct f using the smallest possible number of frame elements. This could be seen as resolving the most important features of f. Note that in the presence of noisy or imperfect data, it may be preferable to seek only an approximate reconstruction. If the frame elements $\{f_i\}_{i=1}^k$ are organized along the columns of an $n \times k$ matrix A, we would thus like the sparsest possible vector of representation coefficients $c \in \mathbb{F}^k$ solving the following problem.

$$\arg\min_c \|c\|_0 \quad \text{subject to} \quad \|Ac - f\|_2 \leq \epsilon. \tag{4.7}$$

Unfortunately, the above ℓ^0-sparse approximation problem has been proved to be

NP-hard [59].

In the face of this intractability, a leading approach is to solve the following convex relaxation of the above problem.

$$\arg\min_{c} \|c\|_1 \quad \text{subject to} \quad \|Ac - f\|_2 \leq \epsilon. \tag{4.8}$$

When $\epsilon = 0$, this is known as the basis pursuit problem, and Chen, Donaho and Saunders [13] show that the associated ℓ^1 minimization often leads to sparse solutions. This additionally holds for the more general case of $\epsilon > 0$, which is described as the basis pursuit de-noise problem; see [11, 22, 79] for further details. The basis pursuit problem can be formulated and solved as a linear program [13]. Scalable methods have additionally been developed to solve the basis pursuit de-noise problem; see, for example, [80]. For additional background on frames and frame-based representations, together with development and application of frame theory in the context of kernel eigenmap methods, including Laplacian Eigenmaps, see [27, 38].

4.3.3 Algorithm Description

With the preceding background established, we present a novel algorithm for constructing a joint embedding of heterogeneous data sets X and Y, derived from distinct and non-overlapping collections of elements. As described at the beginning of the section, our approach is to start with a tentative bijection relating subsets of elements in X and Y. Diffusion maps embeddings are computed using these matched data sets, and then the Coifman-Hirn algorithm-associated operator presented in (4.3) is used to map one embedding into the representation space of the other.

The theory summarized in Subsection 4.3.1 indicates that nearby points in this joint embedding have similar local neighborhood structure. By selecting minimally separated 'mixed pairs', consisting of embedded points derived from data sets X and Y, we can try to align the embeddings around data elements for which the initial bijection is most accurate. We require an adequate set of alignment points, and in particular, we seek a set of pairs for which the corresponding elements in the original data spaces form at least a spanning set or frame. Each element of X and Y can be represented using the data space-specific frame elements, and in keeping with the ideas presented in Subsection 4.3.2, we seek a sparse representation. A final joint embedding is then constructed by mapping each data set element to a frame representation coefficient-weighted linear combination of *jointly embedded* frame elements.

Algorithm Frame-Based Joint Data Embedding

INPUTS:

- Data sets $X \in \mathbb{R}^{N_X \times d_X}$ and $Y \in \mathbb{R}^{N_Y \times d_Y}$.

- Bijection between N-element subsets of X and Y, with $max(d_X, d_Y) \leq N \leq min(N_X, N_Y)$, and corresponding elements specified along the rows of $X' \in \mathbb{R}^{N \times d_X}$ and $Y' \in \mathbb{R}^{N \times d_Y}$.

- Neighborhood sizes: $k_X, k_Y \in \mathbb{N}$.

- Spectral alignment parameters: target value for 2^{nd} eigenvalue $\gamma_2 \in (0, 1)$;

  spectral decay threshold $\gamma_d \in (0, 1)$.

- Minimum frame size: $N_{\mathcal{F}} \geq max(d_X, d_Y)$.

OUTPUT:

- Joint Embedding $Z \in \mathbb{R}^{(N_X+N_Y) \times d}$ of data sets X and Y.

(1) Align Operator Spectra:

- Set σ_X to be the median of the squared distances to the k_X-th nearest neighbor over points in X'; set σ_Y similarly based on Y'.

- Using the matched data sets X' and Y', construct the kernels $K_X, K_Y \in \mathbb{R}^{N \times N}$, with $K_X(i,j) = e^{\frac{-\|x_i - x_j\|^2}{\sigma_X}}$, $K_Y(i,j) = e^{\frac{-\|y_i - y_j\|^2}{\sigma_Y}}$, together with the diagonal matrices $D_X, D_Y \in \mathbb{R}^{N \times N}$ with $D_X(i,i) = \sum_{j=1}^{N} K_X(i,j)$, $D_Y(i,i) = \sum_{j=1}^{N} K_Y(i,j)$.

- Construct the diffusion operators $A_X = D_X^{-1/2} K_X D_X^{-1/2}$, $A_Y = D_Y^{-1/2} K_Y D_Y^{-1/2}$; adjust σ_X, σ_Y if necessary so that the second eigenvalues of A_X, A_Y are approximately equal to γ_2.

- Set the joint embedding dimension to be the smallest integer d such that the eigenvalues $\lambda_X^{(i)}$ of A_X and $\lambda_Y^{(i)}$ of A_Y, ordered by non-increasing magnitude, are less than γ_d for all $i > d$.

Algorithm Frame-Based Joint Data Embedding (continued)

(2) Construct Diffusion Map Embeddings:

- Set $\Lambda_X, \Lambda_Y \in \mathbb{R}^{d\times d}$ to be diagonal matrices containing the d largest eigenvalues, in non-increasing order, of A_X and A_Y, respectively. Let $V_X, V_Y \in \mathbb{R}^{N\times d}$ be matrices containing the corresponding top d eigenvectors along their columns.

- Construct the diffusion map embeddings $M_X, M_Y \in \mathbb{R}^{N\times d}$ of X' and Y', respectively, by computing $M_X = V_X\Lambda_X$, $M_Y = V_Y\Lambda_Y$.

(3) Construct Initial Joint Embedding:

- Apply the Coifman-Hirn algorithm-associated operator specified in (4.3) to map the embedded points in M_X into the embedding space associated with M_Y. In particular, set $M_{X\to Y} = (V_Y^T V_X M_X^T)^T$.

(4) Identify Frames:

- Compute the pairwise Euclidean distances between points specified along the rows of $M_{X\to Y}$ and M_Y in the joint embedding space and rank non-overlapping pairs in order of increasing separation distance.

- Find the smallest $N'_{\mathcal{F}} \geq N_{\mathcal{F}}$ such that the original data space representations of the components of the $N'_{\mathcal{F}}$ nearest embedded pairs span, respectively, $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$.

- Construct the matrices $\mathcal{F}_X \in \mathbb{R}^{d_X\times N'_{\mathcal{F}}}$ and $\mathcal{F}_Y \in \mathbb{R}^{d_Y\times N'_{\mathcal{F}}}$ containing these frame elements for the original data spaces along their columns, and let $Z_{\mathcal{F}_X}, Z_{\mathcal{F}_Y} \in \mathbb{R}^{N'_{\mathcal{F}}\times d}$ denote matrices containing the joint embedding space representations of these frame elements along their rows.

Algorithm Frame-Based Joint Data Embedding (continued)

(5) Construct Final Joint Embedding:

- For $x_i \in X$ (resp. $y_i \in Y$), compute a set of frame representation coefficients with respect to $\mathcal{F}_X$ (resp. $\mathcal{F}_Y$) by solving the sparse representation or approximation problem presented in (4.8). Organize these representation coefficients along the rows of $C_X \in \mathbb{R}^{N_X \times N'_{\mathcal{F}}}$ (resp. $C_Y \in \mathbb{R}^{N_Y \times N'_{\mathcal{F}}}$).

- Compute the final joint embedding by mapping each original space point to an appropriate frame representation coefficient-weighted linear combination of embedded frame elements:

$Z(1 : N_X, :) = C_X Z_{\mathcal{F}_X}$,

$Z((N_X + 1) : (N_X + N_Y), :) = C_Y Z_{\mathcal{F}_Y}$.

Chapter 5

Case Studies in Biomedical Data Analysis

5.1 Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development

5.1.1 Background

Common variations in genetic and epigenetic patterns among humans are associated with variations in risk for developing all common chronic diseases, a few of which have been identified from genome-wide polymorphism screens [10, 86]. The functional biological robustness or its failure in disease is most likely not just reflected in a few dominant components, but in many complex interactions within gene regulatory networks. Due to the overwhelming complexity, the deeper understanding of such networks remains a major challenge in modern systems biology, a field that aims to discover and iteratively refine mechanistic models of biological processes. Biological knowledge is typically encoded in the structure and parameterization of these models. The Gene Ontology project [1, 35] can help to incorporate the known biological details of gene functions into such analysis. The challenge is to reasonably approximate attributes in such models using experimental data that is complex, noisy, and often incomplete.

For the purpose of acquiring biologically rich data sets, laser capture microdis-

section (LCM) has proven a powerful tool to isolate pure cell populations from complex heterogeneous tissue specimens [8, 30, 77]. In combination with microarray technologies, which allow the simultaneous measurement of expression levels for thousands of genes, LCM enables identification of critical gene products even if expressed at low copy numbers.

Our work aims to facilitate efforts in systems biology by organizing data in ways that can suppress noise and better reveal latent, biologically meaningful structure. Coloboma is a not uncommon congenital defect of human ocular development resulting in large retinal holes which often significantly affect vision. The present paper focuses on refinements in the analysis of a temporal series of microarray data obtained from microdissected sites of retinal fissure closure in normal mouse embryos. These data were previously analyzed [9] to identify a putative repressive transcription factor, nlz2 (zinc finger protein 503), which, when its expression was blocked in zebrafish embryos, led to incomplete optic fissure closure, a coloboma model. The interaction of transcription factors, binding sites and gene networks involving nlz2 and related genes, however, are poorly understood [9]. The present paper is dedicated to develop a novel pipeline for the analysis of microarray gene expression data that complements standard approaches and provides a list of candidate genes guiding further experimental analysis of genetic variations.

By developing and applying a novel clustering scheme, we have identified a 50 per cent larger gene cluster (in comparison to PCA and previous hierarchical cluster analyses [9]), whose spatio-temporal gene expressions correlate with nlz2. According to GoMiner, a computational high-throughput tool for biological interpretation of

genomic, transcriptomic, and proteomic data, that identifies the biological processes, functions and components of gene clusters [90, 89], this larger cluster still shows gene enrichment for its specific functions in the context of Gene Ontology.

Next, using GoMiner, we sought to identify those gene clusters whose co-expressions correlate with processes in eye development. First, we apply a novel clustering scheme that builds on the intertwining of Laplacian Eigenmaps, a nonlinear geometrical data transformation, with k-means and hierarchical clustering. To validate the findings, we also use two standard clustering schemes, basic k-means and principal component analysis combined with k-means and hierarchical clustering. All three methods identify gene clusters enriched for functional GoMiner categories related to eye development, but the proposed nonlinear scheme leads to lower false discovery rates. Secondly, we have proposed a mechanism that allows experts to introduce their input in form of additional, labeled information by means of a potential on a data-dependent graph in [20] to improve the dimension reduction and clustering process. Distances between certain labeled genes are forced to appear closer than normally while others are increased. In the present paper, we aim to label genes that are highly connected and thus constitute hubs within the regulatory network. Such genes appear to promote coherence within a gene cluster and would thus be ideal candidates for labeling to obtain a more meaningful and coherent clustering. There are many ways to extract genes of high connectivity, and we use the weights that are generated by the Laplacian on the regulatory network and alternatively weighted correlation networks as described in [49]. Identified gene hubs are then labeled to incorporate regulatory network characteristics into

the labeled Laplacian clustering. This novel clustering scheme based on nonlinear dimension reduction and involving labeled data further improves the biological specificity according to GoMiner analysis. Starting from experimental work based on LCM and microarray technologies in organogenesis, we obtain a list of candidate genes that could be significant in normal development of optic fissure closure and could be useful in guiding analysis of genetic variations in humans with coloboma.

5.1.2 Materials and Methods

The Affymetrix MOE 430 2.0 microarray datasets analyzed to develop and test our new method were for eight samples LCM microdissected from serial cryosections of the retina at the site of final optic fissure closure in the mouse embryos at specific embryonic stages 10.5 days through 12.5 days previously reported in [9]. The 8 time-points span the time just before and just after final fusion (optic fissure closure) and were expected to reveal sets of genes critical for the completion of optic fissure closure in normal development. This previous report further investigated a specific putative repressive transcription factor, nlz2 (or zinc finger protein 503), that was discovered to be highly expressed before and during fissure closure and then downregulated. Gene knockdown experiments in zebra fish of nlz2 resulted in incomplete optic fissure closure (coloboma). Our current analysis explores possible associated gene regulation patterns. Within the 8 different time-point microarrays were 8316 genes consistently identified as expressed and with greater than 2-fold variation in gene expression levels. For our clustering analysis, we chose the subset

of $n = 3416$ genes whose expression levels varied between 4-fold and 26-fold over the 2 days of embryonic development.

For analysis purposes, each gene of the microarray is considered as a vector of its expression levels. This perspective yields a collection of $D = 8$ dimensional vectors. Our proposed analysis relies on Laplacian Eigenmaps [3, 4], a geometrical data transformation that provides a new representation of gene expressions still covering essential geometrical behaviors. The nonlinear geometric representation can be further steered by involving labels [20] that are either derived from weighted correlation networks analysis [49] or from the Laplacian analysis. We intertwine this new data representation with k-means [57], a widely used clustering scheme. GoMiner [90, 89] is then used to identify genes within clusters that are associated with particular biological processes or function.

Let us list the steps of our proposed scheme:

1. **Expression vectors:** Each gene's expression over the 8 time points builds a vector. They constitute a collection $\{x_1, \ldots, x_n\}$ of 8-dimensional vectors, where n is the number of considered expressions

2. **Nonlinear dimension reduction:** Choose a target dimension $d < D$, and obtain a new d-dimensional data representation $\{y_1, \ldots, y_n\}$ of the original D-dimensional vectors $\{x_1, \ldots, x_n\}$

3. **k-means:** Run k-means on $\{y_1, \ldots, y_n\}$ to obtain the final clustering

4. **GoMiner:** Feed the clusters into GoMiner to evaluate their biological relevance

Step 2 in the above scheme is specified in two different ways: First, we use a nonlinear dimension reduction method without labeling (unsupervised):

2.A **Laplacian Eigenmaps:** Choose the number m of gene neighbors and a target dimension $d < D$, then apply Laplacian Eigenmaps to obtain a new d-dimensional data representation $\{y_1, \ldots, y_n\}$ of the original D-dimensional vectors $\{x_1, \ldots, x_n\}$

Alternatively, we may want to incorporate further input into the dimension reduction process by using labeled data. We then identify step 2 with the following supervised procedure:

2.B a) **Identifying highly connected genes:** Apply an R package for weighted correlation network analysis (WGCNA) [49] to identify genes that are highly connected within the gene regulatory network and that act as hubs. Alternatively, use the Laplacian analysis to identify highly connected genes

2.B b) **Schroedinger Eigenmaps:** Gene hubs are labeled by means of a potential term. Choose the number m of gene neighbors and a target dimension $d < D$. The application of Laplacian Eigenmaps with potentials [20] yields a new d-dimensional data representation $\{y_1, \ldots, y_n\}$ of the original D-dimensional vectors $\{x_1, \ldots, x_n\}$

In the following, we present the components of the above scheme in more detail. For comparison we also applied PCA and k-means and therefore briefly discuss these conventional methods too.

Principal component analysis

PCA [61] is a statistical tool that linearly transforms the data into an orthogonal coordinate system whose axes correspond to the principal components in the data, i.e., the first principal component accounts for as much variance in the data as possible and, successively, further components capture the remaining variance. Through an eigenanalysis, the principal components are determined as eigenvectors of the dataset's covariance matrix and the corresponding eigenvalues refer to the variance that is captured within each eigenvector. After subtracting the mean of the dataset, PCA is performed on vectors $\{x_1, \ldots, x_n\}$ by first diagonalizing the covariance matrix $\mathrm{cov}(X) = E(XX^\top)$, where $X = (x_1 \cdots x_n)$ is the zero mean data matrix. The eigenvectors $p_1, \ldots, p_D$ - the principal components ordered according to the magnitude of their eigenvalues - provide the transformed data $Y = W^\top X$, where $W = (p_1 \ldots p_D)$. We obtain the collection of d-dimensional vectors $\{y_1, \ldots, y_n\}$ whose first entries represents the abundance of the primary principal. The second entries are each datapoint's projection along the second eigenvector and so forth.

Laplacian Eigenmaps

Laplacian Eigenmaps (LE) [3, 4] is a nonlinear geometric tool that transforms data into a new representation in a nonlinear fashion. Given points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^D$, we assume that they are steered by d latent variables, and aim to find a new data representation $\{y_1, \ldots, y_n\} \subset \mathbb{R}^d$. We briefly recall the three step procedure of Laplacian Eigenmaps.

Step 1: Adjacency graph, m-nearest neighbors We build a graph $\mathcal{G}$, whose nodes i and j are connected if x_i is among the m-nearest neighbors of x_j or vice versa. The distance between data points is measured by the Euclidean metric. The graph $\mathcal{G}$ represents the connectivity of the data vectors.

Step 2: Heat kernel as weights Next, we weight the edges of the graph and focus on the diffusion *weight matrix* W given by

$$W_{i,j} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\sigma}}, & i \text{ and } j \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

The number of neighbors m controls the sparsity of W.

Step3: Solving an eigenvalue problem We denote a potential new data representation by $y = (y_1, \ldots, y_n)^\top$, where each row is considered as a vector in $\mathbb{R}^d$, and we then consider the following minimization problem

$$\min_{y^\top D y = I} \tfrac{1}{2} \sum_{i,j} \|y_i - y_j\|^2 W_{i,j} = \min_{y^\top D y = I} \operatorname{trace}(y^\top L y), \tag{5.2}$$

where $L = D - W$ and D is the diagonal matrix $D_{i,i} = \sum_j W_{i,j}$. The minimizer of (5.2) is given by the d minimal eigenvalue solutions of $Lx = \lambda D x$ under the constraint $y^\top D y = I$, where I denotes the identity matrix, i.e., the minimizer y's columns are the d eigenvectors with respect to the smallest eigenvalues. If the graph is connected, then $\mathbf{1} = (1, \ldots, 1)^\top$ is the only eigenvector with eigenvalue 0, and we exclude it. Instead of (5.2), we try to find the minimizer of

$$\min_{\substack{y^\top D y = I, \\ y^\top D \mathbf{1} = 0}} \operatorname{trace}(y^\top L y). \tag{5.3}$$

By applying the change of variables $z = D^{1/2}y$, this yields

$$\min_{\substack{z^\top z=I,\\ z^\top \mathbf{1}=0}} \operatorname{trace}(z^\top \mathcal{L} z), \tag{5.4}$$

where $\mathcal{L} = D^{-1/2}LD^{-1/2}$. The minimizer z is given by the d eigenvectors with smallest nonzero eigenvalue, and we obtain the d-dimensional representation $\{y_1, \ldots, y_n\}$ from $y = D^{-1/2}z$.

Identifying highly connected genes

Weighted gene co-expression network analysis is a systems biology tool that allows to identify highly connected genes within a regulatory network. An R package implementation WGCNA is available with an accompanying tutorial [49]. Alternatively, the matrix D in (5.2) is a measure of the connectivity within the network and can be used to identify highly connected genes within the Laplacian framework directly.

Schroedinger Eigenmaps

Based on the Laplacian matrix L in (5.3), a flexible potential, that can capture additional labels, has been introduced in [20]. The matrix L is replaced with a Schroedinger type matrix $E = L + V$, where V is a potential matrix that encodes labels. One then aims to minimize

$$\min_{\substack{y^\top Dy=I,\\ y^\top D\mathbf{1}=0}} \operatorname{trace}(y^\top (L+V)y). \tag{5.5}$$

The result is a new Schroedinger Eigenmaps method that allows for input in an otherwise fully automated dimension reduction process [20]. Here, labels are utilized

to emphasize "important" genes, and we use the connectivity of genes as a measure of their importance in the description of the regulatory network.

Standard cluster analysis

For hierarchical clustering, we refer to [36], and we also apply a shape similarity-based clustering as introduced in [37]. k-means is a method of cluster analysis which aims to partition n observations into k clusters $\{c_1, \ldots, c_k\}$, where k has to be chosen a-priori [57], i.e., one aims at minimizing

$$\arg \min_{c_1,\ldots,c_k} \big(\sum_{j=1}^{k} \sum_{y_i \in c_j} \|y_i - \mathcal{E} c_j\|^2 \big),$$

where $\mathcal{E} c_j$ is the mean of cluster c_j. The basic k-means algorithm requires the target number of clusters to be specified as a parameter.

The k-means algorithm begins with a data set, a target number of clusters k, and a set of $s_1, \ldots, s_k$ initial cluster centroids. It then iteratively assigns points to clusters by centroid proximity, and then adjusts centroids to reflect changes in cluster membership. The algorithm terminates either after a specified number of iterations, or once the cluster centroids/membership no longer change. Although optimal results cannot be guaranteed, the algorithm is quite fast, and many runs can be efficiently computed, with the best clustering taken as an overall result.

GoMiner

GoMiner provides a quantitative and statistical analysis-tool for biological interpretation of genomic, transcriptomic, and proteomic data, commonly derived

from gene expression microarray experiments. It classifies genes into biologically coherent categories and then uses the Gene Ontology project to identify the biological processes, functions and components of genes within these categories [90, 89]. A one-sided Fisher's p-value is used to determine the significance and biological enrichment levels within a category.

Clustering with Genesis

Clustered image maps (CIMs) were first introduced in [83] and are produced here with the Genesis program [76]. We select the Euclidean distance metric and average linkage for hierarchal clustering. To facilitate visualization, a recently-added feature of GoMiner has been implemented that removes large generic categories from all CIMs.

Silhouette coefficient

The silhouette coefficient is a measure for the coherence of clusters. If we take a clustering C to be a mapping from a data set $X = \{x_1, \ldots, x_n\}$ to the integers $1, 2, \ldots, k$ (where k is the total number of clusters), we can define the silhouette coefficient $sil(x)$ for each point x in X to be

$$sil(x) = \frac{B(x) - A(x)}{\max(A(x), B(x))},$$

where $A(x)$ is the average distance between x and other points in its cluster, and $B(x)$ is the average distance between x and the points in the nearest neighboring cluster, cf. [68]. The silhouette coefficient $sil(i)$ for a cluster i is the average of the

coefficients for its constituent points. We similarly define the silhouette coefficient *sil* for an entire clustering to the average silhouette coefficient over all data set points. A clustering with a silhouette coefficient closer to 1 will contain more cohesive and well-separated clusters.

For our experiments, we use the squared Euclidean distance for the computations indicated above, as well as for the data clustering algorithms.

Description of the approach

Microarray data from LCM isolated cells in a mouse model of coloboma as described in the present Section 5.1.2 are analyzed by using standard cluster analysis and a novel gene clustering scheme. We derive a coherent clustering and make use of GoMiner to identify those genes identified in public databases as being associated with eye development or function as a measure of the quality of the other members in the cluster. We used GoMiner to identify the degree of association of clusters obtained by all methods with early stage retinal development, and, in particular, with the closure of the optic fissure, see Figures 1, 2, 3, 4, 5.

For k-means, we set the target number of clusters to be 24, based on previous work with the current data set [9] that yielded biologically meaningful (but smaller and fewer) cluster results. The maximal silhouette coefficient *sil* specifies the best k-means clustering over 100 repeated runs, starting in each case from different randomly selected initial centroids. The maximum was stable over different 100 run sets, suggesting that an at least near optimal clustering was being obtained. Since

the parameter space is too big for an exhaustive search in the dimension reduction process, we fix $\sigma = 1/8$ in (5.1) and assess remaining parameters (number of nearest neighbors and target dimension) over $m = 5, \ldots, 10, 12, 15, 20, 25, 50, 100$ and $d = 1, \ldots, 10, 12, 16$. The idea is that parameter combinations that yield better cluster structure in the mapped data $\{y_1, \ldots, y_n\}$ might be better tuned to resolve possible intrinsic structure in the original data $\{x_1 \ldots, x_n\}$. Silhouette coefficients suggest values $m = 10$ and $d = 2$, which additionally provide excellent GoMiner gene identifications.

5.1.3 Results

We aim to increase our understanding of the gene network underlying the closure of the optic fissure during vertebrate eye development:

Enlarged cluster containing nlz2:

We have identified a 50 per cent larger gene cluster than with hierarchical clustering in [9] whose spatio-temporal gene expressions significantly correlate with nlz2, a gene which when previously inhibited in zebrafish induced coloboma. The latter cluster is associated with 210 Affymetrix probes corresponding to 169 genes, nlz2 was among them. See Figures 6 and 7 for gene expression profiles and its set of enriched functional categories. GoMiner assigns the functional category of 'gene silencing', indicating the repressive influence of nlz2 and co-varying genes. Previous biological studies have shown nlz2 gene product to repress gene transcription of a

number of genes regulated hindbrain development possibly as part of a transcription factor complex consistent with its H2N2 zinc finger domain and its binding site for histone deacetylase. Consistent with this hypothesis, we also identify an additional cluster that varies inversely with the primary 'nlz2 cluster' gene silencing, suggestive of the previously documented role of nlz2 in suppression of gene transcription, cf. Figure 8.

One complementary cluster:

We have found a large cluster whose shape is distinct from nlz2 by applying the similarity-based shape clustering in [37]. GoMiner assigns a number of significantly associated functions to this large cluster including **retina morphogenesis** (vertebrate eye), **generation of neurons**, cellular morphogenesis during differentiation, photoreceptor differentiation, cell motility, **neuron differentiation**, cell projection organization, and biogenesis. The highlighted functions are specifically associated with CHX10, a gene in this cluster that has previously been identified in retinal development, see, for instance, [64, 74].

Collection of enriched clusters:

We also apply k-means on the original data set and on PCA and LE reduced data. The selected 'best' k-means result applied directly to the original data has an overall silhouette coefficient of 0.38. To evaluate PCA+k-means, for each possible number of retained principal components, the mapped data is clustered, and overall silhouette scores are obtained. The best results refer to the mapping based

on principal components capturing about 85% of the variance, with the best overall silhouette score being 0.698. The silhouette scores in the mapped data are substantially higher than those obtained following clustering of the original data, illustrating the fact that Laplacian Eigenmaps enhance cluster structure, see Table 1 for more details.

We find that PCA+k-means, basic k-means, and LE+k-means yield several significantly enriched gene clusters (out of a total of 24) associated with developmental processes, cf. Table 2. Cluster 22 of the Laplacian Eigenmaps-based approach reveals a cluster significantly enriched (with a false discovery rate (FDR) of less than 0.05) for genes specifically implicated in eye development - which is the focus of the experimental work underlying the data set considered in this study. These functional categories (in GoMiner terminology) are

(i) GO:0042462_eye_photoreceptor_cell_development,

(ii) GO:0001754_eye_photoreceptor_cell_differentiation,

(iii) GO:0042461_photoreceptor_cell_development.

When slightly relaxing the FDR up to < 0.15, this cluster 22 shows gene enrichment for further eye specific developmental functions:

(iv) GO:0048592_eye_morphogenesis,

(v) GO:0001654_eye_development,

(vi) GO:0046530_photoreceptor_cell_differentiation,

see also Figures 1 and 4. These categories are neither hit by k-means nor PCA+k-means clustering when restricting the FDR to < 0.05. By relaxing the FDR, however, both k-means and PCA+k-means clustering show gene enrichment for eye specific functions. This verifies that the eye specific functions in LE+k-means cluster 22 are real and have not been picked up by chance. To support the latter claim, we compare the enriched categories in the LE+k-means cluster 22 with the clusters of the other two clustering methods with relaxed FDR. It turns out that specific eye development functions are present in all three clustering methods, but our proposed Laplacian-based scheme leads to lower false discovery rates, see also Table 2. Potential nonlinear structures in the data could be an explanation for this observation, see Figure 9. A nonlinear dimension reduction method would clearly be better suited to fit nonlinear structures than linear methods.

CIMs in Figures 3 a)-b) indicate which clusters across the three methods share common GoMiner categories. It enables us to identify categories that are more specific to one method than to the others. Based on Table 1 the fraction of genes, that are associated to biological functions, are computable for each cluster, method, and false discovery rate.

Note on LE+k-means:

We note that relatively unusual expression patterns are often mapped to distinct, outlying clusters by the Laplacian Eigenmaps approach. For example, the three expression patterns indicated in Figure 8 form a distinct cluster under the

Laplacian Eigenmaps data representation. They are not as well separated in the original and PCA-mapped data, and are consequently misplaced in inappropriate clusters. This could be a technical explanation for greater biological specificity of Laplacian Eigenmaps clustering.

Schroedinger Eigenmaps:

We first label a collection of transcription factors that are known to be annotated to eye development. Enriched GO categories, however, appear generic when applying Schroedinger Eigenmaps with such labels, cf. Figure 10. To obtain more meaningful labels, that are directly extracted from the data rather than from the literature, we identify a set of highly connected genes through the weighted correlation analysis described in [49], see Figures 11 and 12. These "hub genes" are then labeled by means of the potential to steer Schroedinger Eigenmaps utilizing the gene network topology. This labeling seems to further improve the biological specificity, cf. Figure 4. Alternatively, the matrix D in (5.2) is a natural measure of the connectivity within the Laplacian framework. According to D, we use highly connected genes as labels within the LE cluster 22, cf. Figure 13, providing the highest biological specificity, cf. Figure 5. Enriched GO categories that are derived from supervised and unsupervised dimension reduction are shown in Table 3. The supervised procedure Schroedinger Eigenmaps identifies more categories specific to early retinal development and the optic fissure closure than the unsupervised approach.

5.1.4 Discussion

Obtaining a clearer understanding of the gene regulatory network underlying optic fissure closure during eye development will be a long process involving genetic analysis of humans with coloboma and studies of eye development in animal models. Our present analysis and results focus on expanding a list of candidate genes that could be critical for normal fissure closure and in coloboma patients may contain mutations. Compared with conventional clustering algorithms that we tested, our new method is able to identify larger clusters associated either with the nlz2 gene expression or with a distinctly complementary pattern enriched with associations to eye development gene ontologies. It also uniquely identifies the 'nlz2-repressed' pattern as a distinct cluster, cf. Figure 2. The large temporally covarying gene cluster in Figure 7 is identified by GoMiner as being significantly associated with gene silencing, suggestive of a gene regulatory network that represses alternative fates until optic fissure closure is successfully completed (day 11.5 in the mouse). The pattern of genes in Figure 2 could represent such genes that are transiently repressed only when the nlz2 cluster is high. Using temporal pattern-based similarity clustering [37] allows identification of other distinct clusters (i.e., not containing nlz2) in which GoMiner identifies significant associations with specific developmental functions in databases.

Distinct biological specificity for our data set is obtained when labeling highly connected genes and encoding these labels in the potential term. The GO categories **eye morphogenesis**, **retina morphogenesis in camera type eye**, and **camera**

type eye morphogenes, for instance, reflect the optic fissure closure and are identified by Schroedinger Eigenmaps suggesting that nonlinear dimension reduction with labeled data can improve the biological specificity in gene cluster analysis, cf. Figure 5 and Table 3.

Clearly, our new mathematical approach to identify new components of gene regulatory networks controlling development is preliminary and would need further validation to claim its usefulness in more generality. We anticipate improvements in our analysis methods based on nonlinear dimension reduction with connectivity analysis and labeled data.

Microarray data are commonly used for global searches for gene expression changes that might be associated with a perturbation of a cell state or in pathology. In organ development, temporal and spatial patterns accessible through microdissection are associated with reproducible changes in gene expression of even larger numbers of genes. More efficient analysis of microarray data from such microdissected samples could provide improved understanding of cell fate and organogenesis as well as elaboration of gene expression covariance networks. Our nonlinear analysis scheme based on Laplacian Eigenmaps and labeling highly connected genes through a potential appears to offer advantages over standard clustering algorithms in the sense of greater biological specificity and sensitivity. Our results motivate further analysis of nonlinear dimension reduction with labeling within other microarray data sets from LCM dissected tissue or other phenotypically specific cell samples to potentially validate its biological specificity in more generality. Together with LCM-focused gene expression microarray measurements, our proposed analysis

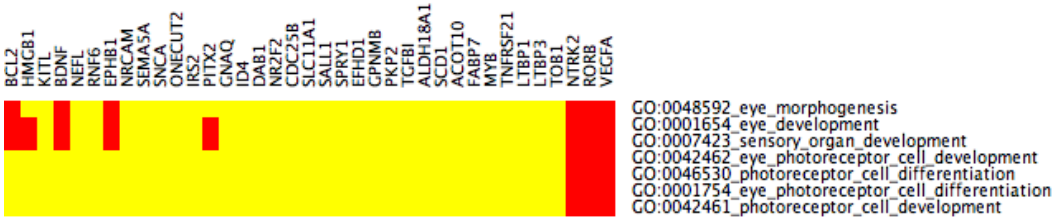
could be part of an iterative process to more completely identify additional elements in gene regulatory networks underlying mammalian organogenesis.

5.1.5 Figures

Figure 1 - CIM cluster 22:

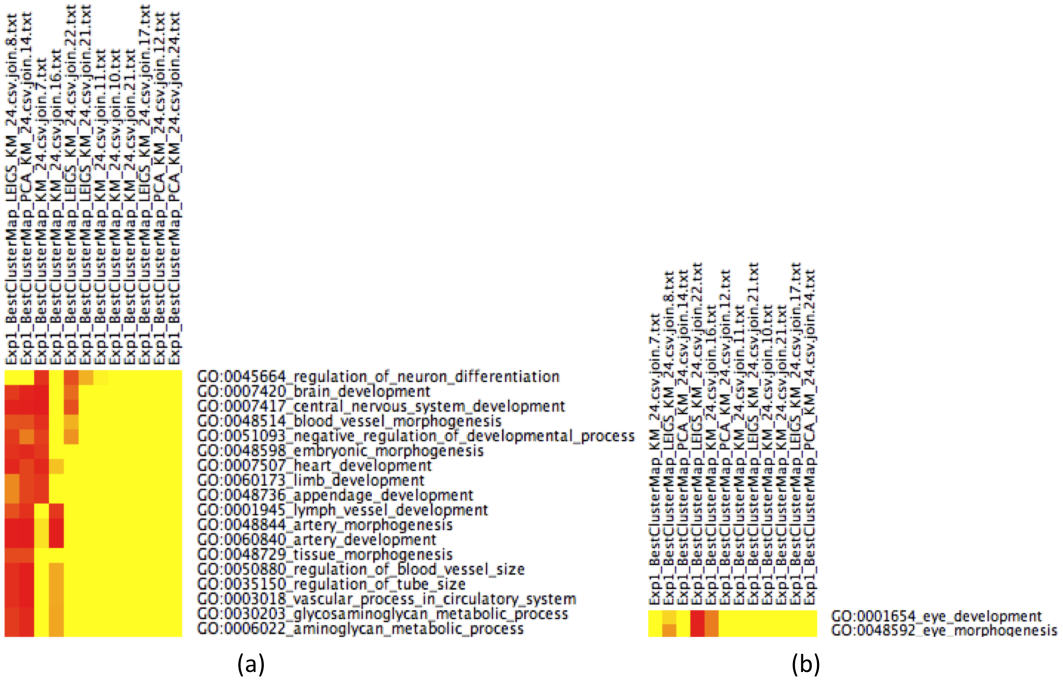
CIM for LE+k-means cluster 22 with functional categories related to eye development, false discovery rate (FDR)$<$ 0.05. The cluster is enriched for eye photoreceptor cell development and for a eye photoreceptor cell differentiation. We hence see GO categories that are closely related to eye development although the FDR is stringently chosen. 24 genes are mapped to 11 GO functions. (Red: genes are mapped to GO categories, Yellow: no association)

Figure 2 - CIM cluster 22 with relaxed FDR:



portion CIM for LE+k-means cluster 22 with functional categories related to eye development (the entire CIM contains 74 GO categories). The input cluster for the present CIM is the same as for Figure 1. By choosing the less stringent FDR< 0.15, more GO categories are statistically enriched, and 36 genes (only 24 in Figure 1) are mapped to these GO categories. Beside the eye related categories in Figure 1, there are additionally eye morphogenesis, eye development, and sensory organ development.

Figure 3 - CIMs across methods:



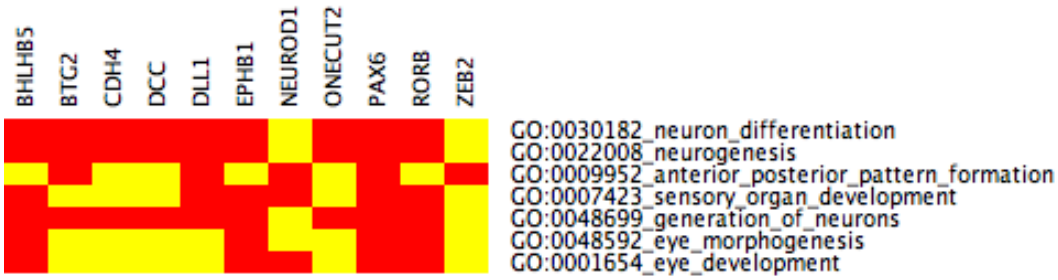
Enriched GoMiner categories that are shared among clustering methods. Each of the 3 different clustering methods (LEIGS KM 24, PCA KM 24, KM 24) produced 24 clusters. We picked 12 among the 72 clusters that seemed to show significant enrichment by means of GO categories. (Yellow means no association. The darker red, the stronger the association between cluster and category) Cluster 22 from Laplacian Eigenmaps+k-means has shown eye related GO categories in Figures 1 and 2 with very stringent FDR. Figures 3 a)-b) verify that these categories have not been picked by chance and that the proposed Laplacian-based scheme leads to lower false discovery rates than the other methods and hence appears to provide greater biological specificity and sensitivity.

(a) GO categories that are shared by at least three clusters, FDR$<$ 0.10. First, cluster 8 of Laplacian Eigenmaps+k-means appears to be closely related to

cluster 14 derived from PCA+k-means. Cluster 22 from Laplacian Eigenmaps+k-means shares few biological functions with cluster 7 of k-means. However, GO categories that are related to eye development are not shared by any other method at FDR$<$ 0.10. Recall that cluster 22 from Laplacian Eigenmaps+k-means has shown enrichment for these categories already at FDR$<$ 0.05 in Figure 1.

(b) The portion CIM with FDR$<$ 0.20 that is associated to additional eye development categories that weren't present in Figure 3a). They are shared by the Laplacian+k-means cluster 22, by k-means cluster 16, and by Laplacian+k-means cluster 8. The entire CIM contains too many GO categories to be listed here.

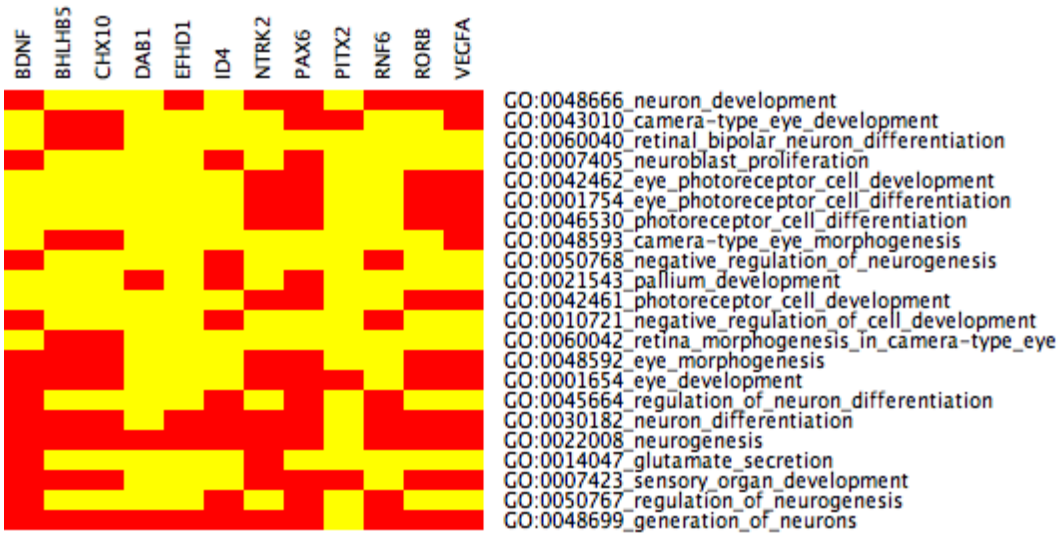
Figure 4 - CIM Schroedinger Eigenmaps II:



Seven highly connected genes from Figures 11 and 12 were labeled in Schroedinger Eigenmaps. After clustering, all seven labeled genes are contained in the same cluster with 145 other genes. The cluster is enriched for categories (eye morphogenesis, eye development) that are more specific to eye development than the results without labeling suggesting that data-dependent gene labeling can increase the biological

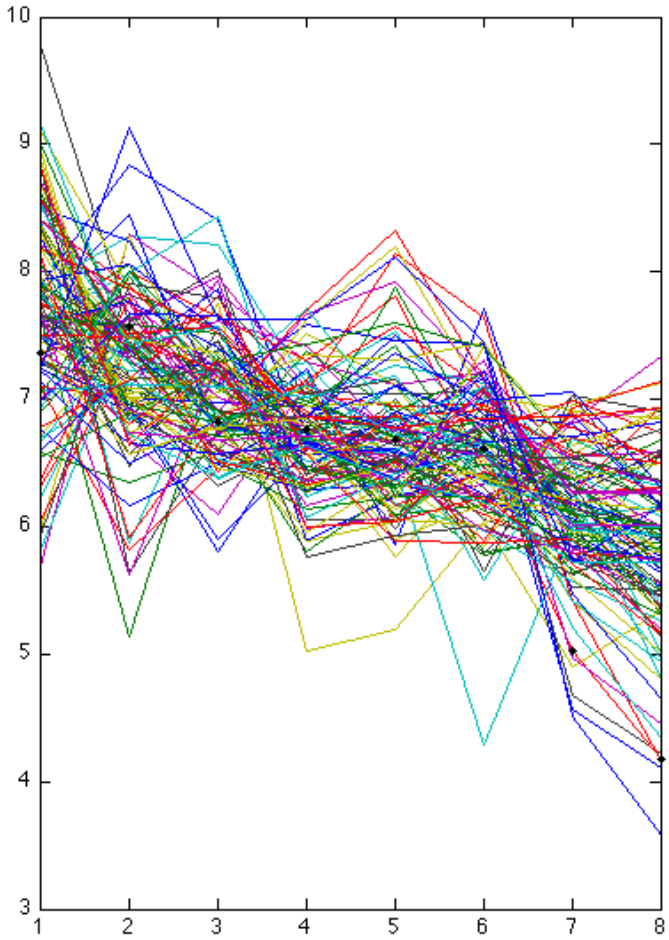
specificity.

Figure 5 - CIM Schroedinger Eigenmaps III:



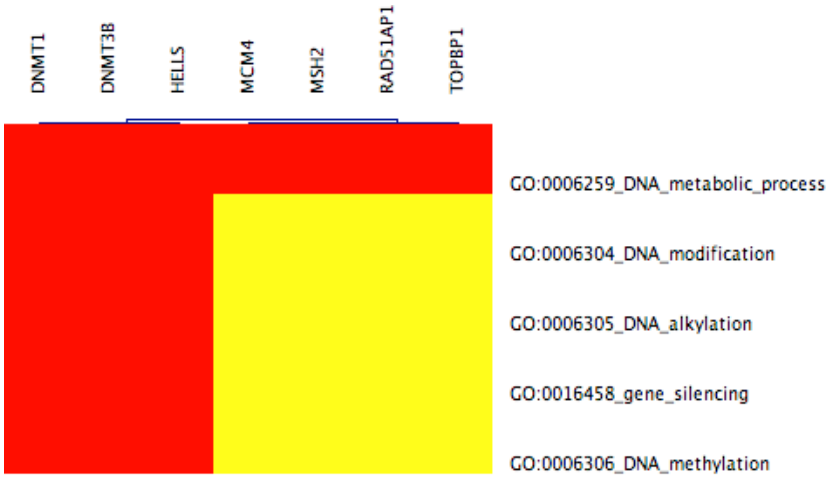
Five highly connected genes from Figure 13 were labeled in Schroedinger Eigenmaps. After clustering, all five genes are contained in the same cluster. The GO categories for this cluster match the optic fissure closure and thus provide distinct biological specificity (eye photoreceptor cell development, eye photoreceptor cell differentiation, camera type eye morphogenesis, retina morphogenesis in camera type eye, eye morphogenesis, eye development, sensory organ development). Schroedinger Eigenmaps using labels derived from the Laplacian weight matrix D provide better specificity than using Laplacian Eigenmaps without any labels.

Figure 6 - nlz2 cluster profile:



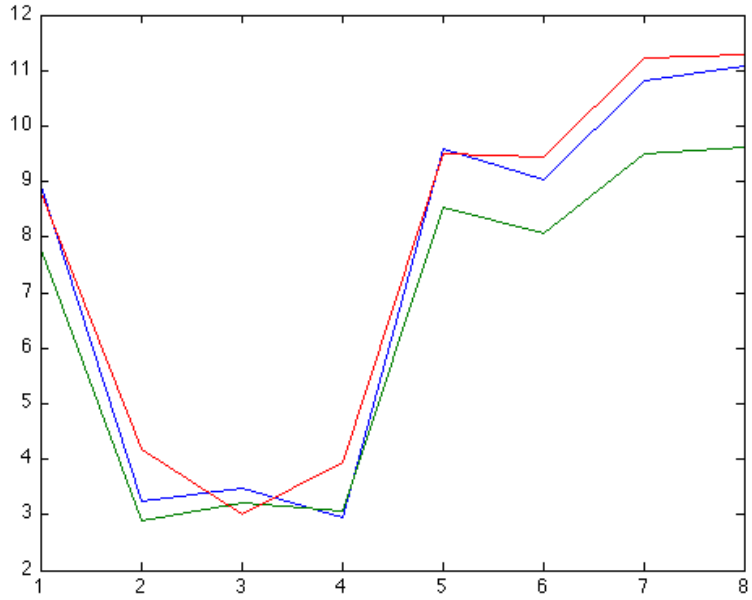
Profile of cluster that contains nlz2. Gene expression levels are plotted vs. 8 time-points, black circles indicate nlz2.

Figure 7 - CIM containing nlz2:



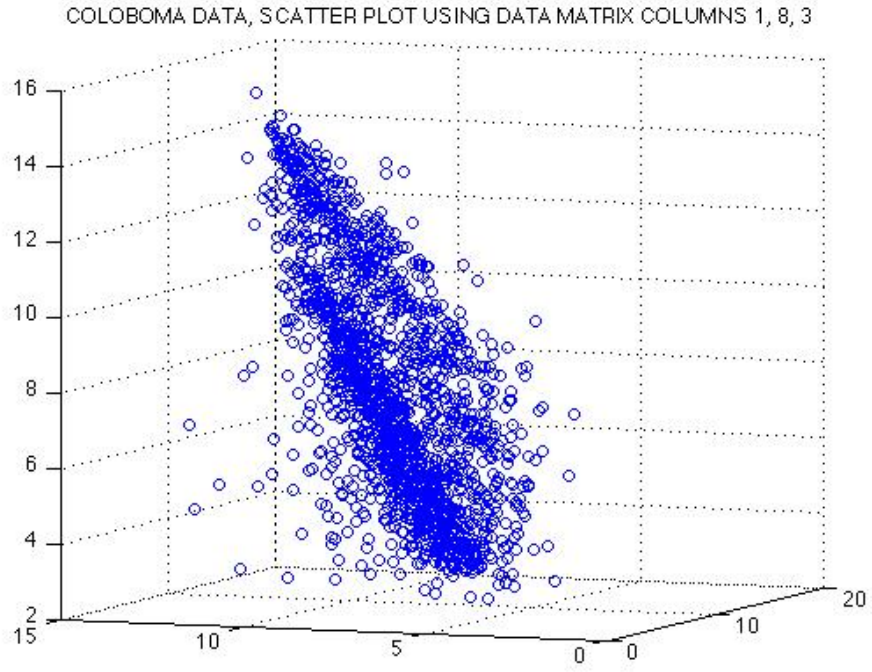
Clustered Image Map (produced by GoMiner) showing enriched functional categories for the cluster that contains nlz2 and 168 other genes. More genes in this cluster have been associated to the 5 above GO categories (gene silencing is among them) than one would expect by chance. The 7 genes (DNMT1,...,TOPBP1) above are mapped to these GO categories within the GoMiner database. Red indicates that genes were mapped to GO categories. Yellow means no annotation. Due to gene expression co-variation within the cluster, other genes in the cluster could possibly related to the above GO categories too. Since gene silencing is associated to this cluster, one may speculate that nlz2 and co-varying genes have repressive function and that there is a cluster that shows the reverse expression profile, see Figure 8.

Figure 8 - outliers:



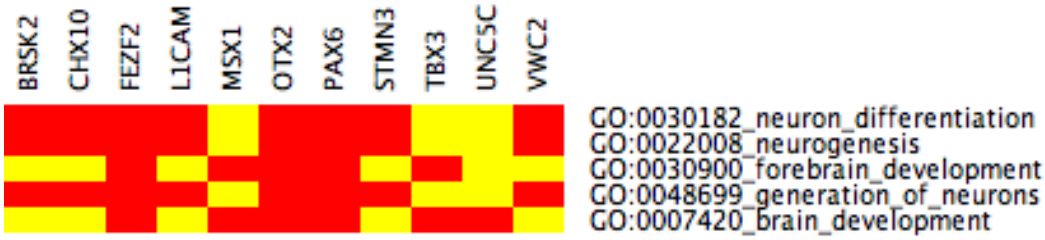
Outliers that LE+k-means captures into a separate cluster, the associated Affymetrix probes are 1427262_at, 1427263_at, 1436936_s_at. All three probes are associated to XIST, a gene that is transcribed and spliced but does not appear to encode a protein. XIST inactivation is known to be an early developmental process in mammalian females.

Figure 9 - The 8-dimensional data are projected onto a 3-dimensional subspace:



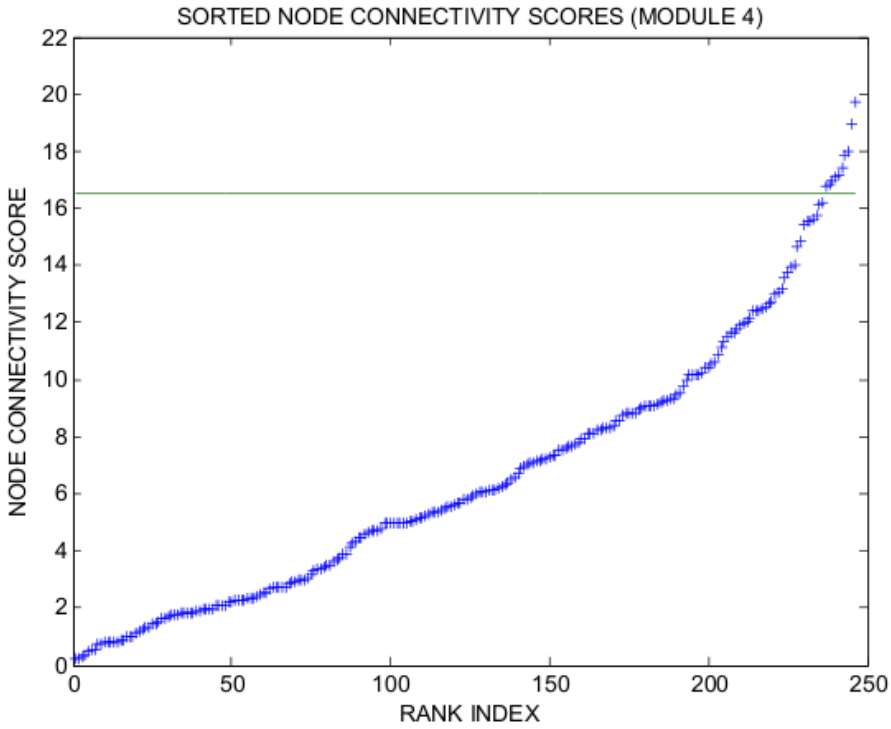
The 3-dimensional subspace is spanned by their 1st, 3rd, and 8th coordinates. If the data would lie on a linear subspace in $\mathbb{R}^8$, then the projected data must show a linear pattern. However, the actual projection of our data does not show a linear pattern but rather two cones next to each other. A nonlinear approach like Laplacian Eigenmaps could be useful to recover nonlinear structure of the data manifold.

Figure 10 - CIM Schroedinger Eigenmaps I:



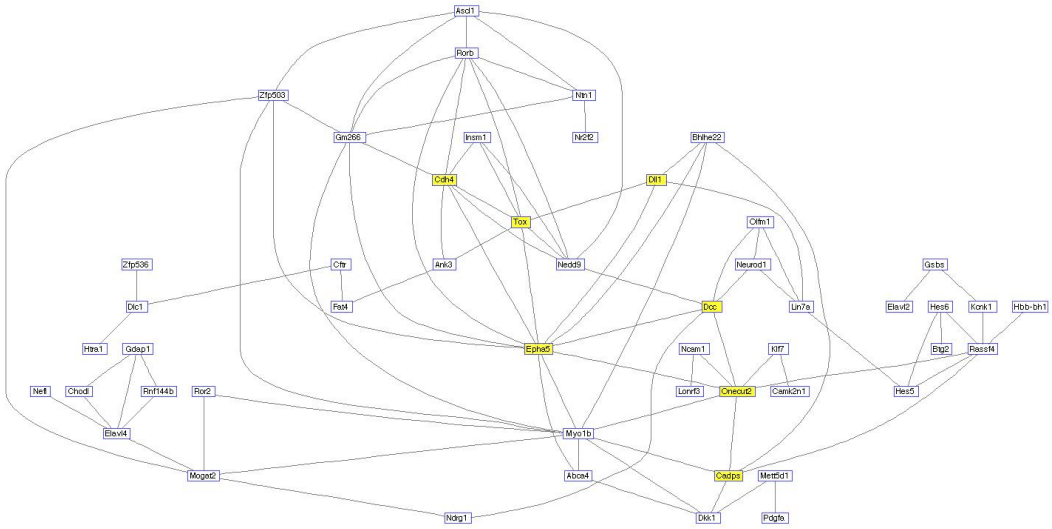
We have labeled transcription factors (CHX10, OTX2, PAX6) that are known from the literature to be associated to development. We intend to derive a cluster whose GO categories are specific to eye development when starting with labeled genes. Schroedinger Eigenmaps using these labels is applied and the new data representation is then clustered. The three labeled TF are contained in the same cluster with 154 other genes. GoMiner enrichment analysis leads to relatively generic categories that one would expect from the choice of the labeled TF. However, specific eye development categories were not found. This observation suggests that gene labeling based on literature search leads to relatively poor GoMiner enrichments.

Figure 11 - plot of connectivity scores in increasing order for WGCNA weights:



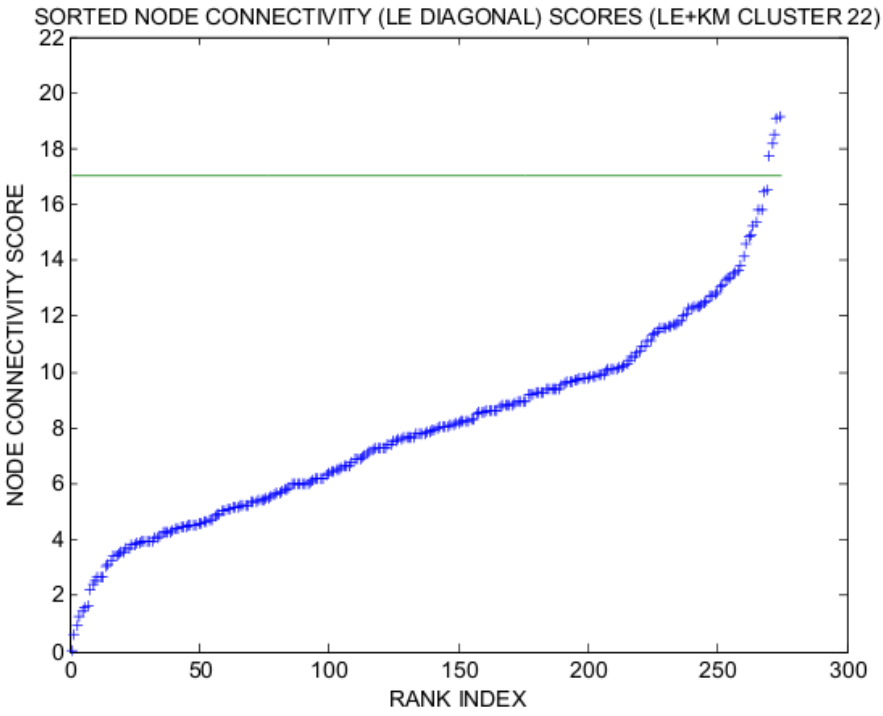
To derive gene labels directly from the measured Affymetrix data rather than from the literature, we aim to identify co-varying genes with high connectivity in the regulatory network. The plot shows the connectivity of Affymetrix probes within a cluster enriched for eye development computed by WGCNA. Rank index refers to Affymetrix probes. The associated most highly connected genes according to WGCNA are Cdh4, Dll1, Tox, Onecut2, Dcc, Epha5, Cadps.

Figure 12 - connectivity network for WGCNA weights:



Portion of the thresholded weighted correlation network derived from WGCNA. In the entire connectivity network, each of the genes to be labeled (Cdh4, Dll1, Tox, Onecut2, Dcc, Epha5, Cadps) would have more than 16 connections, see also Figure 11.

Figure 13 - plot of connectivity scores in increasing order for LE weights, LE cluster 22:



We aim to further improve the biological specificity of cluster 22 derived from Laplacian Eigenmaps + k-means. To identify co-varying genes with high connectivity in the regulatory network of cluster 22, we measure connectivity by means of the weight matrix D in (5.2). The connectivity of Affymetrix probes within LE+KM cluster 22 are shown. The most highly connected genes in cluster 22 are Etv3, Zfp386, Kdm4c, Eea1, Fyttd1, which can be used as labels in Schroedinger Eigenmaps.

5.1.6 Tables

Table 5.1 comparison for unsupervised methods: Silhouette coefficients and number of genes for each cluster and unsupervised clustering method (no labels). Laplacian Eigenmaps+k-means leads to higher silhouette coefficients.

	k-means		PCA+k-means		LE+k-means	
cluster	sil	# genes	sil	# genes	sil	# genes
1	0.0200	65	0.7329	126	0.6535	103
2	0.3067	146	0.6221	60	0.7049	125
3	0.4078	180	0.7002	168	0.6862	174
4	0.4068	234	0.6840	198	0.6848	154
5	0.3401	255	0.7423	157	0.7831	97
6	0.2960	252	0.7033	130	0.7949	389
7	0.3442	90	0.6795	126	0.7369	120
8	0.6509	9	0.6800	65	0.6953	270
9	0.3900	254	0.6393	190	0.7800	91
10	0.2162	34	0.7130	187	0.7046	79
11	0.3056	112	0.6517	182	0.7606	141
12	0.3531	165	0.7162	155	0.7487	122
13	0.4636	182	0.6925	117	0.9889	3
14	0.4267	167	0.7422	205	0.7118	125
15	0.6529	114	0.6968	184	0.5997	85
16	0.1593	86	0.5266	9	0.7214	236
17	0.5488	13	0.6792	84	0.6839	83
18	0.4323	253	0.6956	211	0.7380	135
19	0.1749	20	0.7151	118	0.6466	72
20	0.3076	133	0.6926	170	0.7243	121
21	0.4314	174	0.7041	115	0.7461	199
22	0.4394	130	0.7342	116	0.7442	275
23	0.4538	210	0.7252	192	0.6849	115
24	0.4366	138	0.6792	151	0.8534	102

Table 5.2 The number of enriched Go-categories are counted over all 24 clusters at a false discovery rate of 0.05 which is the default configuration of GoMiner. k-means and PCA+k-means do not show any eye specific enrichement in any of the clusters. Only LE+k-means provides one cluster that is enriched for 3 categories specific to eye development. These categories would have even been picked at an FDR of 0.01 suggesting strong statistical support for the LE+k-means performance. Potential nonlinear structures in the data could be an explanation for this observation, see Figure 13.

	k-means	PCA+k-means	LE+k-means
# enriched Go-categories	55	17	27
# enriched Go-categories specific to eye development	0	0	3

Table 5.3 comparison between supervised and unsupervised methods: GO categories related to the optic fissure closure that are associated to clusters derived from unsupervised (no data labels) and supervised (labeled data) methods. Using labels that are computed directly from the measured data appears to provide more biological meaningful associations than unsupervised methods.

unsupervised methods	Schroedinger Eigenmaps + WGCNA / *D*-labels
eye morphogenesis	
eye development	
eye photoreceptor cell development	
eye photoreceptor cell differentiation	
photoreceptor cell development	
embryonic morphogenesis	
morphogenesis of a branching structure	
sensory organ development	
	camera type eye morphogenesis
	retina morphogenesis in camera type eye
	retinal bipolar neuron differentiation

5.2 Predicting expression-related features of chromosomal domain organization with network-structured analysis of gene expression and chromosomal location

5.2.1 Introduction

A growing range of experimental results indicate that the cell nucleus is highly organized on many levels [55]. Chromosomes, while not set in fixed locations, do appear to form relatively stable associations within particular territories. Smaller, gene rich chromosomes (chromosomes 16, 17, 19, 20, 21, and 22) often aggregate toward the interior of the nucleus, while chromosomes with reduced gene density, as well as gene-free telomeric regions, tend to be distributed around the nuclear periphery [55]. Individual chromosomes are themselves variably complexed and compacted through hierarchically structured interactions with histones and other DNA-binding proteins. The resulting chromatin landscapes shape the relative accessibility of genes, and thus, many facets of their expression. An interesting prospect is that gene expression and chromosomal organization may be substantially related, with relatively stable patterns of inter-chromosomal interactions emerging in support of coordinated expression programs [46, 28, 62]. This sort of organization could allow transcription factor networks to be locally structured and more dynamically responsive. For example, a factor acting on genes dispersed over several chromosomes could function more efficiently if the relevant regions are juxtaposed, allowing smaller concentrations to quickly diffuse over a restricted nuclear neighborhood.

This paper presents a flexible approach for detecting features of chromosomal domain organization that may be related to coordinated gene expression. The essential idea is to identify chromosomal neighborhoods over which genes are relatively co-expressed. These locally correlated clusters (LCCs) are then associated, yielding a candidate, expression-related inter-chromosomal interaction network. For the key step of identifying LCCs, we apply a non-linear data representation approach - Laplacian Eigenmaps (LE) - to organize genes with respect to a combined measure of co-expression and physical proximity along a chromosome.

Previous studies have demonstrated co-expression domains at various genomic scales, with neighboring genes being more likely to exhibit co-expression than more distant ones, even after accounting for clusters of duplicated genes [85, 16, 53]. Additional tools have been developed to partition and visualize microarray data with respect to chromosomal location [87]. While these efforts motivate our work with valuable insights, their approach and emphasis is somewhat different. In particular, the preceding studies directly analyze chromosomal location-structured correlation maps to identify contiguous co-expression domains, and more global patterns of co-expression at different genomic scales. Our aim is to present a somewhat more specific framework for detecting the most prominent locally correlated domains that are additionally correlated across chromosomes. The selection approach for these co-expression domains is potentially more sensitive, in that the LE-based data organization allows detection of locally correlated clusters that need not include every gene along a chromosomal segment. The overall output is a network of potential expression-related inter-chromosomal interactions that is amenable to detailed anal-

ysis. This expression-based chromosomal interaction network view is motivated by recent modeling studies, which integrate expression data with matching data on physical interactions between chromosomes [62].

In the following sections, we describe the Laplacian Eigenmaps technique, as well as the general approach used to identify and relate LCCs. We then present some results with a gene expression data set derived using 5 state-of-the-art microarray platforms over the widely-studied NCI-60 cancer cell lines. In particular, we show that numerous LCCs can be identified and inter-related. The resulting candidate inter-chromosomal interaction network features a prominent hub cluster, with two known transcription factors. We describe two levels of network validation, the first statistical, and the second with respect to experimentally measured chromosomal interactions in a published study. A particular strength of the current study is the breadth and comprehensiveness of the NCI-60 derived data set, which is among the most detailed available for a system of comparable biological complexity. We conclude by discussing some ongoing work to further evaluate and extend the presented methods, taking advantage of the range and structure of the available profiling data.

5.2.2 Data Sets

Gene transcript expression was derived for the NCI-60 cancer cell lines using five leading microarray platforms.: the Affymetrix (Affymetrix Inc., Sunnyvale, CA) Human Genome U95 Set (HG-U95) (GEO accession number GSE5949) [71]; the Human Genome U133 (HG-U133) (GEO accession number GSE5720) [71]; the

Human Genome U133 Plus 2.0 Arrays (HG-U133 Plus 2.0) (GEO accession number GSE32474) [65]; the GeneChip Human Exon 1.0 ST array (GH Exon 1.0 ST) (GEO accession number GSE29682) [65] and the Agilent (Agilent Technologies, Inc., Santa Clara, CA) Whole Human Genome Oligo Microarray (WHG), (GEO accession number GSE29288) [56]. For these gene expression data sets (accessible at `http://discover.nci.nih.gov/cellminer/`) all probes were put through rigorous quality control. The first criterion was for each accepted probe expression profile (across the 60 cell lines) to have an intensity range r satisfying $\log_2 r > 1.2$. The second criterion was to use probes with a minimum average correlation to all related probes of 0.60 when possible, or of 0.30 ($p < 0.02$) if not. The probe expression profiles that passed these steps were then standardized and averaged to obtain gene-specific expression profiles integrating data derived from the various platforms.

5.2.3 Laplacian Eigenmaps for Nonlinear Data Organization

We apply a nonlinear data organization technique - Laplacian Eigenmaps (LE) - to organize genes with respect to a combined measure of co-expression and physical proximity along a chromosome. Each gene can be represented by a set of measurements - expression values across cell lines, as well as a chromosomal position. The measurements can be seen as points in a Euclidean space, and we can reasonably suppose that these points - particularly the expression values - are somewhat constrained by the highly structured networks of underlying molecular interactions. In more mathematical terms, we assume that our data set consists of points $\{x_1, ..., x_N\}$

drawn from a d-dimensional manifold in $\mathbb{R}^D$ ($d << D$), or a distribution with support concentrated on a d-dimensional manifold. We obtain a manifold structure preserving low dimensional representation $\{y_1, ..., y_N\} \subset \mathbb{R}^d$ in three steps:

1. Construct Data Adjacency Matrix W: for $k \in \mathbb{N}$, put an edge between elements i and j if x_i is among the k nearest neighbors of x_j or vice versa. Weight connected edges using $W_{ij} = e^{\frac{\|x_i - x_j\|^2}{\sigma}}$, $\sigma > 0$.

2. Construct Laplacian Matrix L: Set $D_{ii} = \sum_j W_{ij}$, and let $L = D - W$.

3. Compute Eigenmaps: Solve $Lx = \lambda Dx$. Let $f_0, f_1, ..., f_d$ be the eigenvectors corresponding to the first $d + 1$ eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_d$. Discard f_0 and embed in d-dimensional space using the map $x_i \mapsto y_i = (f_1(i), f_2(i), ..., f_d(i))$.

The overall approach is motivated by the fact that the graph-based Laplacian matrix L can be seen as a discrete analogue of the Laplace-Beltrami operator on the underlying manifold. The eigenmaps of the latter operator provide an optimal embedding of the manifold into a space of reduced, intrinsic dimension. Since the graph-based Laplacian converges to the manifold-based Laplace-Beltrami operator, its associated data mappings progressively inherit the corresponding manifold recovery guarantees [4, 5].

We can also understand LE in more concrete terms. Let the $N \times d$ matrix $y = (y_1, ..., y_N)^T$ denote the low-dimensional representation of our data set. The above

eigenvalue problem $Lx = \lambda Dx$ can be shown to solve the following minimization:

$$\underset{(y^T Dy=I)}{\arg\min}\, \mathrm{trace}(y^T Ly) = \underset{(y^T Dy=I)}{\arg\min}\, \frac{1}{2} \sum_{i,j} \|y_i - y_j\|^2 W_{i,j}. \tag{5.6}$$

Note that the first term on the right forces neighboring points in the original data space (with large W_{ij}) to be mapped next to one another. As such, LE acts to organize data with respect to local features, which allows natural cluster structure to be better revealed. LE can, in fact, be related to spectral clustering techniques which approximate optimal graph partitionings [4]. LE can additionally be shown to be equivalent to kernel PCA using a kernel matrix associated with the commute times of diffusion on the underlying data graph [33]. These commute times are compressed in regions of natural cluster structure, where many paths connect given pairs of points. By preserving commute time distances, the LE-based representation thus preserves cluster structure and reflects aggregate connectivity relationships between data set elements, as captured by the matrix W. The mathematical integration of information over many paths relating network-structured data elements also tends to suppress noise-related features in the data. Taken together, the above attributes often allow LE to enhance established learning approaches by providing a more meaningful representation of the data. In a gene expression study of embryonic eye development, we have shown that LE can be combined with standard clustering techniques to identify sets of genes with significantly increased functional specificity [25, 26, 88]. Other studies have shown that LE can be applied to improve classification and visualization of gene expression microarray data [2, 52]

5.2.4 Fusion of Gene Expression and Chromosomal Location Data

Information on gene expression and chromosomal location was combined within the framework of LE. In particular, we applied the following steps for each chromosome.

- Construct a symmetric, k-nearest-neighbor matrix W_{loc} relating genes with respect to base pair distances

  ($W_{loc}(ij) \neq 0$ if gene i is among the k nearest neighbors of gene j or vice versa).

  - Set k to be the smallest value yielding a connected data graph.
  - Weight distances using a Gaussian kernel, i.e., $W_{loc}(i,j) = e^{-\frac{(bpdist(i,j))^2}{\sigma}}$, where $bpdist(i,j) > 0$ denotes the separation in base pairs between genes i and j, and σ is set to give genes separated by 2Mbp a weight of $exp(-\frac{1}{2})$.

- Construct a symmetric expression similarity matrix W_{exp}, with values specified according to the 'mask' implied by W_{loc}:

  - set $W_{exp}(i,j) = 1$, if $W_{loc}(i,j) \neq 0$ *and*

    AbsoluteCorrelation$(i,j) > 0.3$; (for the study data set, an absolute correlation of 0.3 was estimated to be associated with a p-value just under 0.05, making all selected correlations significant with respect to this threshold).

  - set $W_{exp}(i,j) = 0$, otherwise.

- Perform Laplacian Eigenmaps using $W = W_{loc} + W_{exp}$, followed by k-means clustering of the mapped data.

The idea is to perform LE with respect to a symmetric ($W_{ij} = W_{ji}$) 'fusion kernel' W, which strongly weights genes that are both co-located and co-expressed. Symmetry, together with the nonnegative entries of W, assure that the constructed Laplacian matrix L is positive semidefinite and diagonalizable, yielding well-defined eigenmaps as described in the previous section [4]. The location kernel (W_{loc}) bandwidth σ is set to impose a 'soft' local neighborhood threshold around 2Mbp. The target dimensionality for the LE mapping and the number of clusters selected for k-means clustering were both set to the same value, based on the 'spectral gap' associated with the Laplacian matrix derived from W. K-means clustering was performed 100 times, with the best clustering selected based on average silhouette score (a measure of intra-cluster cohesion and inter-cluster separation). The use of the spectral gap for selecting the target dimensionality and cluster count is motivated by a natural connection between LE and spectral clustering methods that also operate from the graph Laplacian L. Basic results from spectral graph theory show that the eigenspace of eigenvalue 0 for L is spanned by indicator vectors $\mathbb{1}_{C_1}, ...\mathbb{1}_{C_k}$ corresponding to the connected components $C_1, ...C_k$ of the data graph [15, 82]. In the scheme presented above, the LE nearest neighbor parameter is selected to yield a connected data graph, but any intrinsic cluster structure in the data will produce strongly connected subcomponents that are only weakly connected to one another. In this setting, the eigenvectors selected with respect to the spectral gap (e.g., with

eigenvalues closer to zero) correspond to approximate indicator vectors for these natural clusters in the data graph.

The presented approach for fusing gene expression and chromosomal location data is one of several possibilities in a general LE/kernel-based framework. Some alternatives are indicated in the Discussion section. For this initial study, we made specific choices to accentuate, through the LE mapping, even modestly correlated clusters that are well localized. The fusion kernel is thus structured with respect to the location data graph relating genes and gene neighborhoods along a chromosome. The LE nearest neighbor parameter is set to assure a connected location data graph, so that all genes are related, while the LE kernel bandwidth parameter is set to extract reasonable local neighborhoods in the data mapping. The expression kernel is binary-valued, with pairwise expression values set to 1 whenever a genes are within a neighborhood and their correlation exceeds a statistical significance-based threshold. The idea here is to integrate expression-based relationships between genes, as might be derived from a significance-thresholded correlation network. An alternative is to apply exact absolute correlation values in the expression kernel, but we found that this often causes relatively separated genes that are very highly correlated to be strongly weighted in the fusion kernel (and thus mapped next to one another). More sophisticated parameter tuning could perhaps address this issue, but the indicated approach was applied for this study to place the focus on identifying and relating a set of well-defined LCCs.

We finally note that the use of the Gaussian kernel for weighting location distances, as well as the additive combination of location and expression kernels,

is motivated by the theory supporting Laplacian Eigenmaps. The essential idea is that the Laplace-Beltrami operator is intimately related to heat diffusion on manifolds, with diffusion processes integrating and reflecting local manifold geometry. In this sense, the Laplace-Beltrami operator is a fundamental geometric object, and its eigenfunctions naturally support data transformations that preserve manifold geometry [5]. The action of the Laplace-Beltrami operator on a differentiable function defined on a manifold can be expressed in terms of the heat kernel, which in the appropriate local coordinate system is approximately the Gaussian. The Gaussian kernel thus arises naturally in the discrete approximation to the Laplace-Beltrami operator applied in Laplacian Eigenmaps [4, 67]. Additive combination of kernels fits well within this mathematical framework, while alternatives such as point-wise multiplication of kernels are somewhat more removed. With the latter, we have nonzero kernel weights of the form $e^{(-\|\cdot\|_{loc}^2 - \|\cdot\|_{exp}^2)}$, where $\|\cdot\|_{loc}$ and $\|\cdot\|_{exp}$ indicate pairwise location and expression measures, respectively. The resulting argument to the exponential is no longer even the value of a metric, which breaks the connection to the described, diffusion-based approximation framework. We have developed related techniques for analysis of joint data-dependent graphs and their associated diffusion kernels in the context of hyperspectral imagery. These approaches, fusing spatial and spectral information, have produced the best known classification results with several classical data sets [6, 7].

5.2.5 Results

After applying the LE-based data fusion approach presented above, followed by k-means clustering, sets of genes that were both co-located and co-expressed were identified. These locally correlated clusters (LCCs) were k-means-based clusters for which (1) all genes were situated within a 4Mbp neighborhood, and (2) the average pairwise absolute correlation relating cluster genes was greater than 0.2. Candidate inter-chromosomal interactions were then identified by selecting LCCs on different chromosomes with average pairwise (inter-cluster) absolute correlation greater than 0.18. These thresholds were motivated by the relative strength of the significance-based pairwise absolute correlation threshold for the study data (0.3), as well as the aim of obtaining an interaction network of reasonable size, but still amenable to detailed analysis. Empirical significance estimates are provided for the resulting intra and inter chromosomal interactions using an approach detailed in the results section.

Applying the procedure detailed above, we were able to identify 114 LCCs, distributed over 17 chromosomes. Their distribution is presented in Figure 5.1. Only two LCCs (57 and 58) overlap, though many are situated fairly close to one another. As previously noted, the average pairwise absolute correlation between cluster genes was applied as a co-expression measure, and in what follows, we will refer to this as the *intra-cluster similarity measure*. The maximum value for the latter over the 114 LCCs was 0.4993, with a median value of 0.2306. To estimate the relative significance of these values, we applied a re-sampling-based approach. Specifically,

for each LCC, we constructed 10000 random sets of genes (drawn from the entire data set), with each set taken to be the same size as the LCC (in terms of gene count). We then computed the intra-cluster similarity measure for each of these random sets. Finally, we derived a p-value for each LCC's observed intra-cluster similarity by taking the fraction of random gene sets with greater than or equal intra-cluster similarity. For the majority of LCCs (103), no random set yielded equal or greater intra-cluster similarity, indicating a p-value < 0.0001. For the remainder, the maximum estimated p-value was 0.0009.

To assess interactions between LCCs, we once again computed the average pairwise absolute correlation, but now with respect to 'mixed pairs', i.e. with one gene taken from each of the two LCCs. We will refer to this average absolute correlation as the *inter-cluster similarity measure*. Applying a threshold of 0.18, we obtained a network (shown in Figure 5.2) of 60 LCCs participating in a total of 87 inter-chromosomal interactions. To estimate the significance of these interactions, we applied a re-sampling-based approach similar to the one used for assessing the intra-cluster similarity for individual LCCs. For each pair of LCCs, we computed the inter-cluster similarity for 10000 randomly selected gene set pairs (with the same gene count sizes as the relevant LCCs), and derived a p-value by taking the fraction exceeding the particular observed inter-LCC similarity. For 22 of the 87 inter-LCC interactions, no random sets yielded equal or greater inter-cluster similarity, indicating a p-value < 0.0001. For the rest, the maximum estimated p-value was 0.0007.

The interaction network shown in Figure 5.2 features a clear hub cluster, LCC

101. The latter is situated on chromosome 22, and prominently interacts with numerous LCCs on chromosome 19. These candidate interactions are consistent with the general observation that chromosomes 19 and 22 are among a group of small, gene-rich chromosomes known to preferentially interact with one another [55]. To further assess the obtained results, we compared the proposed LCC interactions with published chromosomal contact maps obtained using the Hi-C technique [55]. In Figure 5.3, some contact map segments corresponding to the neighborhoods associated with particular LCC 101 hub interactions (shown in Figure 5.2) are highlighted, and appear to occur in high contact frequency regions. One caveat is that these physical interaction maps were derived from a human cell line distinct from the NCI-60 set used in this study. Still, one perspective is that prominent, expression-related features of chromosomal domain organization may be relatively conserved over a range of human cell types and physiological conditions.

Three notable features of the computed LCC interaction network are (1) its dominant hub LCC 101, which makes 25 connections with other chromosomal domains, (2) the relatively few predicted inter-chromosomal interactions associated with other LCCs directly connected to the dominant hub, and (3) the fairly substantial number of interacting elements relative to the total set of LCCs (60 of 114) . To quantify and combine these attributes, we consider the following multi-component measure computed from an adjacency matrix A:

$$\|A\| = \frac{1}{deg(TopHub)} + conn(TopHubCC) + \frac{1}{\|spec(A)\|_1}.$$

Here $deg(TopHub)$ indicates the number of connections associated with the top hub;

$conn(TopHubCC)$ is the relative connectivity of the connected component of the top hub, i.e., the fraction of observed edges relative to all possible edges; $\|spec(A)\|_1$ is the L^1 norm of the set of eigenvalues of A, which is related to its rank, and increases with the number of nodes involved in interactions. The essential interpretation is that substantial, hub-structured networks will have smaller values of the above measure, which is constructed from the most elementary, relevant mathematical measures. Using this measure, we tried to assess the significance of the overall computed network structure featuring the dominant hub LCC 101. To do so, we constructed 1000 sets of 114 LCC length-matched random intervals, with at least 15 genes per random interval. (The latter restriction gives a median gene count comparable to that observed over the 114 actual LCCs.) Adjacency matrices were derived for each of these random networks by thresholding the inter-chromosomal interactions at the similarity (average absolute correlation) value of 0.18, just as with the constructed LCC interaction network. We found that only 41 of the 1000 random interaction networks had a smaller value for the above multi-component measure than the LCC interaction network, giving the latter an estimated p-value of 0.041.

5.2.6 Discussion

When considering the hub-structured LCC interaction network of Figure 5.2 in the context of gene expression-related features of chromosomal domain organization, an immediate question is the gene content of the main hub cluster 101. Its

Table 5.4 Genes associated with hub LCC 101.

ANKRD54 BAIAP2L2 C22orf23 CARD10
CDC42EP EIF3L GCAT GGA1 H1F0 LGALS1
MAFFtf MFNG MICALL1 PDXP PICK1
PLA2G6 POLR2F SH3BP1 SOX10tf TRIOBP

20 associated genes are listed in Table 5.4. We note here that two transcription factors are included in the localized cluster. One, MAFF, is a basic leucine zipper transcription factor that is known to bind to an element in the promoter of the oxytocin receptor (OTR) gene. The MAFF gene may also function more broadly to regulate elements of the cellular stress response, though specific targets in LCCs interacting with LCC 101 do not seem apparent. A second transcription factor in LCC 101 is SOX10, a member of the of the SOX (SRY-related HMG-box) family of transcription factors regulating aspects of cell fate determination during embryonic development. Relatively few direct targets of the SOX10 gene product have been identified, with 11 known or suspected targets discussed in [54]. Of these, only one possible target, NGFR, was found in an LCC (61), but the latter was not among the LCCs identified as inter-chromosomal interaction candidates. Overall, the dominant hub-structured network of Figure 5.2 does not seem to be obviously explained by specific, known regulatory interactions.

We additionally considered the set of genes contained in LCC 101 and its immediately connected nodes. If these LCCs indeed represent a set of chromosomal domains interacting in support coordinated gene expression, we would expect some

enrichment for functionally related genes. To assess this, we applied the High-Throughput GoMiner tool to perform a gene set enrichment analysis within the framework of the Gene Ontology [89]. We found that the 852 genes associated with the LCC 101 connected component were significantly enriched ($p < 0.01, FDR < 0.01$) for genes associated with immune responses. These included some known gene clusters (e.g., a Human Leukocyte Antigen group on chromosome 6) as well as additional genes drawn from multiple interacting LCCs.

More generally, we note that the presented approach for identifying and relating LCCs should be seen within a broader framework, built on kernel-based data fusion and nonlinear data organization. This particular study is something of a proof-of-concept, with certain somewhat strong assumptions made with the aim of obtaining reasonable initial results for further assessment. For example, the applied 'fusion kernel' was somewhat location driven, with neighborhood (LE nearest neighbor) parameters set to yield a connected location data graph that effectively related all genes along a chromosome. This may be a natural choice for many gene-dense chromosomes, whose location data graphs could be connected with small nearest neighbor parameter settings. For larger chromosomes, with sparser or less uniform gene distributions, a reasonable alternative approach might be a more modest parameter setting that yields a multi-component location data graph. Another LE parameter - the location kernel bandwidth σ - could also be set on a chromosome by chromosome basis. The latter was specified to impose a soft threshold around a 2Mbp local neighborhood, but refinements might adjust this based on chromosome size and gene density. Our approach also integrated pairwise co-expression

data with respect to the location-structured data graph, i.e., with the expression kernel recording values only for neighborhood genes. The latter kernel was also binary-valued, with pairwise expression similarities set to 1 whenever a statistical significance-based correlation threshold was exceeded. The intention with all these choices was to accentuate, through the LE mapping, even modestly correlated clusters that are well localized. An approach beginning with a non-binary-valued expression kernel is also possible. For an initial study, we avoided this because the parameter selection is somewhat less clear. Without careful tuning, the many highly correlated but still (linearly) dispersed genes can come to dominate the combined location and expression kernel.

We generally note that the presented framework is very flexible, and we hope to assess variations building on its essential strengths:

- kernel-based fusion of heterogeneous data types
- nonlinear data organization in support of clustering and other learning approaches
- network-structured interaction maps amenable to detailed analysis and statistical validation.

We are currently working to analyze relatively homogeneous subsets of the NCI-60 (e.g., epithelial and melanoma-derived cell lines) to consider how the identified LCC interaction networks might be altered, perhaps toward more tissue-specific gene groups. Although the present focus has been on inter-chromosomal interaction networks, construction of intra-chromosomal interaction networks is possible with

the developed methods. These can potentially be focused toward noteworthy regions initially identified through analysis restricted to inter-chromosomal interactions. We are additionally working to develop theoretically-motivated refinements to kernel construction and combination.

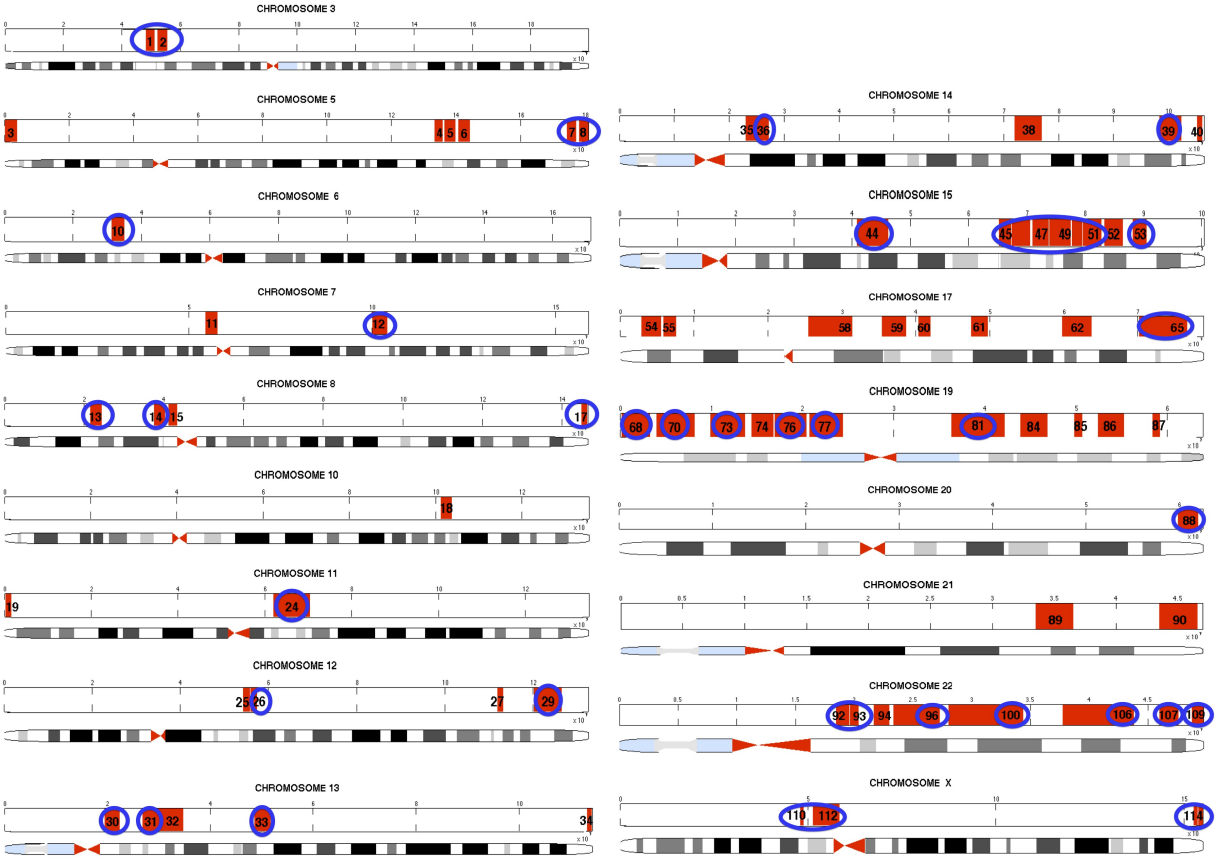


Figure 5.1: Locations of locally correlated clusters. Circled clusters are additionally correlated with other clusters on different chromosomes. Clusters generally do not overlap (with just one exception), though some are situated relatively close to one another. Clusters are numbered sequentially with respect to chromosome and start location. In marked regions containing two or more closely spaced clusters, the region is numbered with respect to the last associated cluster.

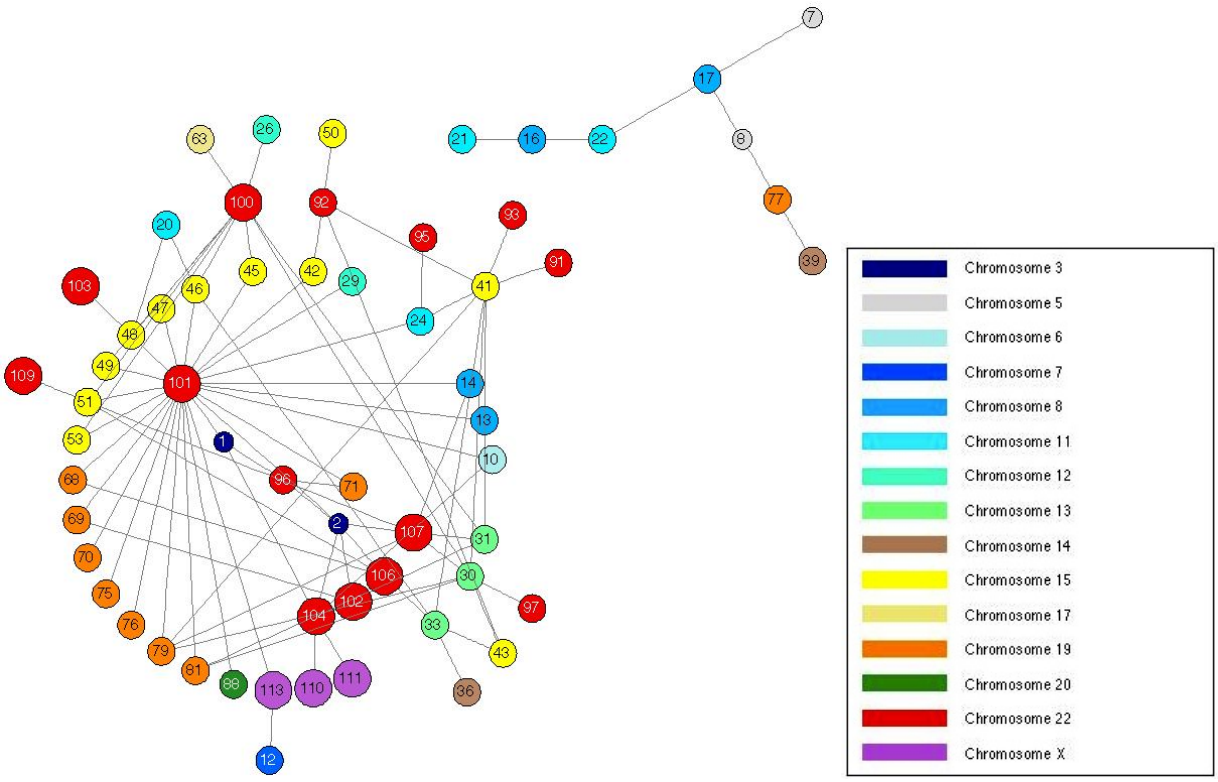


Figure 5.2: Inter-chromosomal association network for locally correlated clusters. (Numbering as in Figure 1, with colors used to indicate distinct chromosomes.)

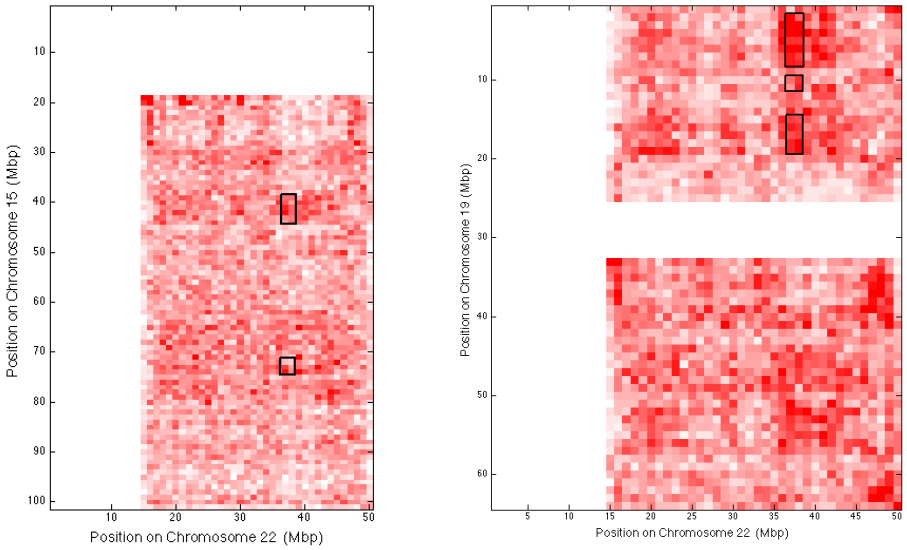


Figure 5.3: LCC 101 is associated with several other LCCs on different chromosomes, as indicated by the clear hub structure seen in the interaction network of Figure 2. Some of these expression-based interactions are consistent with experimentally measured physical interactions between human chromosomal regions. Lieberman-Aiden et al. applied the Hi-C method to obtain contact maps between 1-Mbp regions on different chromosomes in the human lymphoblastoid cell line GM06690 [55]. The intensity of each pixel in the above maps indicates the contact frequency between two such regions. Overall, several of the computed candidate interactions between the hub LCC 101 and LCCs on chromosomes 15 and 19 match regions that have been determined to interact experimentally. These results support the idea that co-expression of gene clusters on different chromosomes may be facilitated by relatively stable patterns of chromosomal organization that place relevant regions in close proximity [46, 28, 62]

LCC	LCC	Hi-C data p-value
17	7	0.0027892
17	8	0.00048508
21	16	0.0033632
22	16	0.0033632
22	17	0.0022422
101	1	0.0016457
101	2	0.0074085
101	10	0.0048493
101	42	0.00494
101	68	0.0077963
101	69	0.0015593
101	70	0.0056054
104	1	0.00092407
104	10	0.001248
107	31	0.0057768

Figure 5.4: Each of the LCC interaction-associated Hi-C data blocks shown in Figure 3 can be associated with a p-value. This is computed as the fraction of (equal size) data blocks in the pairwise inter-chromosomal interaction map with contact frequency equal or greater than that of the observed block. The p-values for the blocks in the upper left map are 0.0054 (top) and 0.1214 (bottom). In the upper right map, no other equal-size blocks matched or exceeded the contact frequency associated with the top block. The middle and bottom blocks had p-values 0.0967 and 0.0111, respectively. P-values can similarly be computed for each of the 87 pairwise LCC interactions shown in Figure 2. Applying the Benjamini-Hochberg procedure for controlling the false discovery rate (FDR) associated with a family of hypothesis tests, the 15 LCC interactions indicated in the left table were determined to be significant at FDR = 0.05.

5.3 Identification of Drug Response-Related Gene Sets

5.3.1 Introduction

For most anti-cancer drugs, relatively little is known about the detailed mechanism of action. Even where targets have been defined, as with FDA-approved and in-clinical-trial drugs, broader off-target effects remain poorly understood. These notably include polypharmacology, as well as the integrative pathways beyond initial targets that ultimately determine efficacy [40]. Cancer emerges though genetic and epigenetic alterations that perturb molecular networks controlling cell growth, survival, and differentiation [34, 81]. To develop more targeted and efficacious cancer treatments, it is essential to situate and understand drug actions in this networked, systems-level context. The National Cancer Institute's Laboratory of Molecular Pharmacology (NCI-LMP) has developed a rich and growing array of databases to support this foundational objective [66]. These include drug compound chemoactivity data over the widely studied NCI-60 cancer cell lines, together with detailed molecular profiling data for each cell line, such as transcript expression, gene copy number, and gene sequence.

In this study, we present the results of an integrated analysis of gene expression and chemoactivity profiling data over the NCI-60 cell lines. In particular, we use a nonlinear data representation technique, Laplacian Eigenmaps [4], to organize the chemoactivity data and identify coherent clusters of compounds sharing similar response profiles over the NCI-60. The clusters are then organized in a network based on their relative similarity. This drug cluster network is highly concordant with the

existing understanding of compound class relationships, grouping known mechanism of action drugs into coherent clusters, together with novel compounds sharing similar response profiles. At the same time, the drug cluster network, organizing over 20000 compounds, reveals numerous clusters of response profile-related drugs with little or no relation to clusters enriched for known mechanism of action drugs. These drug compound clusters may represent agents that are active against novel targets and pathways.

To better understand the gene sets and pathways implicated in their responses, we merge the drug cluster network with a set of gene co-expression network modules inferred using the WCGNA algorithm [41] applied to baseline gene expression data over the NCI-60. This approach yields a joint network of drug clusters and gene co-expression modules, relating groups of drugs to sets of genes whose expression profiles are highly correlated with their response profiles. We show that sets of known mechanism of action drug compounds lie within clusters that are linked to co-expression network modules containing genes known to be implicated in their response. We also find a novel interaction between a coherent drug cluster with no known mechanism of action drugs and a co-expression network module organized around the gene encoding the Bmx non-receptor tyrosine kinase. The Bmx protein product has been implicated in several types of cancer, and is emerging as a highly attractive target for drug development [39, 44].

We conclude by presenting some initial results from a novel approach to integrated analysis of the gene expression and drug compound chemoactivity profiles. While both of these types of data are ostensibly vectors of measurements over the

NCI-60, their intra-class correlation structure is notably different. In particular, the drug compound activity profiles have a distribution of pairwise correlations that is notably shifted toward positive values. As a consequence, if compound activity profiles and gene expression profiles are directly clustered together, most compounds fall into clusters with few to no genes, in spite of significant compound-gene associations. To try and organize the compounds and genes according to such relations while preserving intra-group structure, we apply the joint embedding algorithm presented in Subsection 4.3.3. The results show some favorable attributes, which we present, while discussing remaining challenges.

5.3.2 Data Sets

The analysis results presented in Subsection 5.3.4 are based on the data sets described below.

- G is a 26065×60 matrix with gene expression profiles over the NCI-60 organized along the rows. 10498 expression profiles have a single missing value in one of three cell lines.

- C is a 20602×60 matrix with drug compound chemoactivity profiles over the NCI-60 organized along the rows. All chemoactivity profiles have observations for at least 35 cell lines. 18012 profiles have at least one missing value. Among these profiles, the median number of missing values is 4.

- G' is a 26065×57 matrix of complete gene expression profiles derived from G by excluding 3 cell lines associated with all missing values.

- C' is a 4149×57 matrix of complete drug compound chemoactivity profiles derived from C, over the same set of 57 cell lines with gene expression measurements recorded in G'.

We use e.g., $G(i, \cdot)$, to denote the i^{th} gene expression profile in G, with the corresponding notation similarly used to indicate profiles derived from the other data sets. We additionally use N_G, N_C, $N_{G'}$, and $N_{C'}$ to denote the number of drugs or compounds in the data sets G, C, G', and C', respectively.

The platforms used to provide the gene expression data are the Affymetrix (Affymetrix Inc., Sunnyvale, CA) $\sim 60,000$ feature Human Genome U95 Set (HG-U95) [71]; the $\sim 44,000$ feature Human Genome U133 (HG- U133); the $\sim 47,000$ feature Human Genome U133 Plus 2.0 Arrays (HG-U133 Plus 2.0), the $\sim 5,500,000$ feature GeneChip Human Exon 1.0 ST array (GH Exon 1.0 ST) [65], and the Agilent (Agilent Technologies, Inc., Santa Clara, CA) $\sim 41,000$ feature Whole Human Genome Oligo Microarray [56]. For these gene expression data sets (accessible at `http://discover.nci.nih.gov/cellminer/`) [66], all probes were put through rigorous quality control. The first criterion was for each accepted probe expression profile (across the 60 cell lines) to have an intensity range r satisfying $\log_2 r > 1.2$. The second criterion was to use probes with a minimum average correlation to all related probes of 0.60 when possible, or of 0.30 ($p < 0.02$) if not. The probe expression profiles that passed these steps were then standardized and averaged to obtain gene-specific expression profiles integrating data derived from the various platforms.

We use the term 'drug' to indicate chemical compounds tested in the NCI-60 Developmental Therapeutics Program (DTP) human tumor cell line screen [73]. The latter screen uses the NCI-60 cancer cell lines to prioritize novel compounds showing selective growth inhibition of particular tumor cell lines. These cell lines encompass 9 tissues of origin, including breast, central nervous system, colon, lung, prostate and renal cancers, as well as leukemia and melanomas. To assess potential associations between drug activity and mRNA expression levels, we use the 50% growth inhibitory concentrations (GI50) determined by the DTP. The activity levels are specifically expressed as the negative log of the 50% growth inhibitory concentration [-log10(GI50)], measured using a 48-hour sulphorhodamine B assay. The drug activity data set include 353 drugs with putatively known mechanism of action. Drug activity profiles are standardized in a manner analogous to the gene expression profiles.

5.3.3 Analysis Methods

Drug Cluster - Gene Co-Expression Module Network Construction

The following procedure was applied to construct a network of drug compound clusters and gene co-expression modules.

1. Compute the $N_C \times N_C$ drug compound activity profile similarity matrix S_C, with $S_C(i,j) = Cor(C(i,\cdot), C(j,\cdot))$, where Cor denotes the Pearson's Correlation computed over matched, non-missing entries in $C(i,\cdot)$ and $C(j,\cdot)$.

2. Construct the $N_C \times N_C$ compound data adjacency matrix W_C, with $W_C(i,j) = e^{\frac{-(1-Cor(C(i,\cdot),C(j,\cdot))^2}{\sigma_C}}$,

   if $C(i,\cdot)$ is one of the k nearest neighbors of $C(j,\cdot)$ by 'correlation distance' $(1 - Cor(C(i,\cdot), C(j,\cdot))$, or vice versa; set $W_C(i,j) = 0$ otherwise.

   Select k to be the smallest value such that the data graph represented by W_C is connected ($k = 16$ for data set C).

   Set σ_C to be the median squared correlation distance to the k^{th} nearest neighbor over all N_C drug compounds.

3. Apply Laplacian Eigenmaps, essentially described in Section 2.4, but starting from the application-specific construction of the data adjacency matrix W_C described in the preceding step. Let Y_C denote the $N_C \times d_C$ matrix used to record the embedded data. ($d_C = 119$, based on observation of a corresponding 'gap' in the ordered spectrum of the Laplacian matrix L_C constructed from W_C).

4. Cluster the embedded compound activity data recorded along the rows Y_C using average linkage hierarchical clustering with Euclidean distances. Apply the Dynamic Tree Cut algorithm [50] to derive a set of clusters from the hierarchical cluster tree. Let $\mathcal{C}_C$ denote the set of drug compound clusters.

5. Apply the WCGNA algorithm [41] to construct a gene co-expression network and a corresponding set of co-expression modules using pairwise absolute correlation values. Let $\mathcal{M}_G$ denote the set of co-expression modules.

6. For each compound cluster in $\mathcal{C}_C$, identify the *cluster hub compound* as the compound with the strongest total, i.e., summed, pairwise correlation to other cluster compounds. Similarly, identify the *module hub gene* for each co-expression module in $\mathcal{M}_G$.

7. Form a drug cluster - gene co-expression module network by linking elements of $\mathcal{C}_C$ and $\mathcal{M}_G$ if their hub-to-hub correlation values exceed a threshold $\tau > 0$.

Drug Compound Activity Profile - Gene Expression Profile Joint Embedding

The joint embedding algorithm introduced in Subsection 4.3.3 was applied to the data sets G' and C'. In particular, in terms of the notation associated with the algorithm description, we set $X = C'$ and $Y = G'$. The joint embedding was thus constructed by mapping embedded compound data into the embedded gene data space. An initial bijection between subsets of $X = C'$ and $Y = G'$ was constructed by setting $X' = C'$ and $Y' = G'_{cor}$, where G'_{cor} was a subset of gene expression profiles in G' matched to the set of compound activity profiles in C' based on the Pearson's correlation. In particular, pairwise correlations between the activity profiles in C' and the expression profiles in G' were computed, ranked in decreasing order, and used to select the strongest correlation-based gene-compound pairings. The algorithm parameters were set to be $k_X = k_Y = 16$, $\gamma_2 = 0.95$, $\gamma_d = 0.1$, and $N_{\mathcal{F}} = 200$. These settings yielded a joint embedding dimension of $d = 240$. We will refer to the jointly embedded compound activity and gene expression data sets as

$C_J \in \mathcal{R}^{N_{C'} \times d}$ and $G_J \in \mathcal{R}^{N_{G'} \times d}$, respectively.

5.3.4 Results and Discussion

Applying the initial analysis workflow presented at the start of Subsection 5.3.3, we obtain a network of 105 drug compound clusters shown in Figure 5.5. Known mechanism of action compound classes are distributed within clusters in a manner consistent with understanding of their relationships. For example, DNA damaging agents, such as Topoisomerase 1 and 2 inhibitors and alkylating agents are concentrated in a clique of tightly interacting clusters in the upper left corner of Figure 5.5. Some cluster statistics for these highly connected DNA-damaging agent-enriched clusters are presented in a table above the set of clusters. The cluster coherence is indicated by the median correlation of cluster compounds to the cluster hub compound, together with the corresponding interquartile range (IQR). Classes of compounds, such as the kinase inhibitors, known to act on a broader range of protein targets and pathways show a more distributed cluster distribution.

Figure 5.6 shows the same set of 105 clusters shown in Figure 5.5, with additional links indicating strong interactions with particular co-expression network modules. A correlation threshold of 0.60 is used for ease of visualization in the figures, though correlations greater than 0.26 are significant at $p < 0.05$ for the study data. To assess the extent to which known compound-response related gene interactions are reflected in the drug cluster - co-expression module network, we examined linked clusters and modules for these underlying elements. Some examples

of recovered associations are shown for DNA damaging agents and kinase inhibitors in Figures 5.7 and 5.8, respectively.

Figure 5.6 indicates a substantial number drug clusters with no known mechanism of action compounds. To investigate these, we ranked such clusters by their coherence, taking the ratio of the median hub correlation to the corresponding IQR, and selected the most coherent clusters participating in strong interactions with similarly coherent co-expression modules. As an example, we mention one such interaction, shown in Figure 5.9. The implicated drug compound cluster, D40, is highly coherent, with a median hub correlation of 0.84, and corresponding IQR of 0.29. Its interacting co-expression module is organized around the Bmx gene, which, as noted in the opening discussion, is a non-receptor tyrosine kinase implicated in several cancers. A pattern comparison of the Bmx gene expression and D40 hub compound activity profiles shows that the strong association is driven by a single, in this case, melanoma cell line. We have verified the accuracy of the associated measurements at the level of the raw gene expression and compound activity data. Although these strong associations do not guarantee that the D40 hub compound is targeting Bmx, as compared to perhaps some other co-regulated protein in its signaling pathway, there are no other cancer-implicated kinases or similarly likely targets in the Bmx module. We are currently evaluating compounds in the D40 drug cluster for possible assessment against Bmx in a direct kinase inhibition assay.

The strong interactions indicated in Figure 5.6, which include only a fraction of the statistically significant interactions, show that the gene and drug data graphs can be meaningfully inter-related. An immediate idea is to directly cluster the gene

expression and compound activity profiles, as measurements over a common set of cell lines. The limitations of such a straightforward approach are illustrated in Figures 5.10 and 5.11. Figure 5.10 shows that the genes and drug profiles have rather different pairwise correlation distributions, with the compounds showing a pronounced shift toward positive correlations. This is not surprising in view of the development of the drug databases under study, with many compounds synthesized through small variations of existing ones, leading to groups of compounds with highly correlated activity profiles. As consequence, a direct clustering of compound and gene profiles places most compounds in clusters with few genes. This is illustrated in the left plot of Figure 5.11. A more detailed accounting of the cluster composition shows that, with a direct clustering using the k-means algorithm, just over 2/3 of the compounds fall into clusters with a fraction of compounds greater than 90%.

To try and organize the compounds and drugs in a manner consistent with the significant interactions shown in Figures 5.6, 5.7, and 5.8, we applied the joint embedding algorithm presented in Subsection 4.3.3, as described in Subsection 5.3.3 of the present study. As with the original gene and compound profiles, k-means was applied to the jointly embedded gene and drug data. The initial results of these experiments were mixed. Compounds were less concentrated in compound-specific clusters, as compared with the clustering based on the original data. The cluster compositions associated with the jointly embedded data are shown in the right plot of Figure 5.11. With the jointly embedded data, just under 1/5 of the compounds were in clusters with a fraction of compounds greater than 90%. Some clusters with substantial numbers of activity profile-correlated, known mechanism of action

compounds were found, showing that the embedding preserved certain expected relations. But overall, the correlations between compounds and genes co-embedded and clustered together was not strong as with seen with the cluster network - gene co-expression module analysis.

Consideration of the theory developed in Section 4.3 suggests that the root problem could be the quality of the initial bijection constructed as described in Subsection 5.3.3, based on correlations between a set of compounds and a much larger set of genes. We are experimenting with alternative approaches for 'seeding' the initial alignment of the data networks. One promising idea is to cluster the gene and compound data separately, identify cluster hubs, construct a bijection between these hubs, and use this bijection to run the joint embedding algorithm of Subsection 4.3.3. This approach is similar in spirit to the construction of drug cluster - gene co-expression module network illustrated in Figure 5.6. It seems plausible, with a reasonable number of clusters, since the joint embedding algorithm only requires enough cluster hub-specified points to form frames, i.e., spanning sets, for the respective compound and gene data spaces.

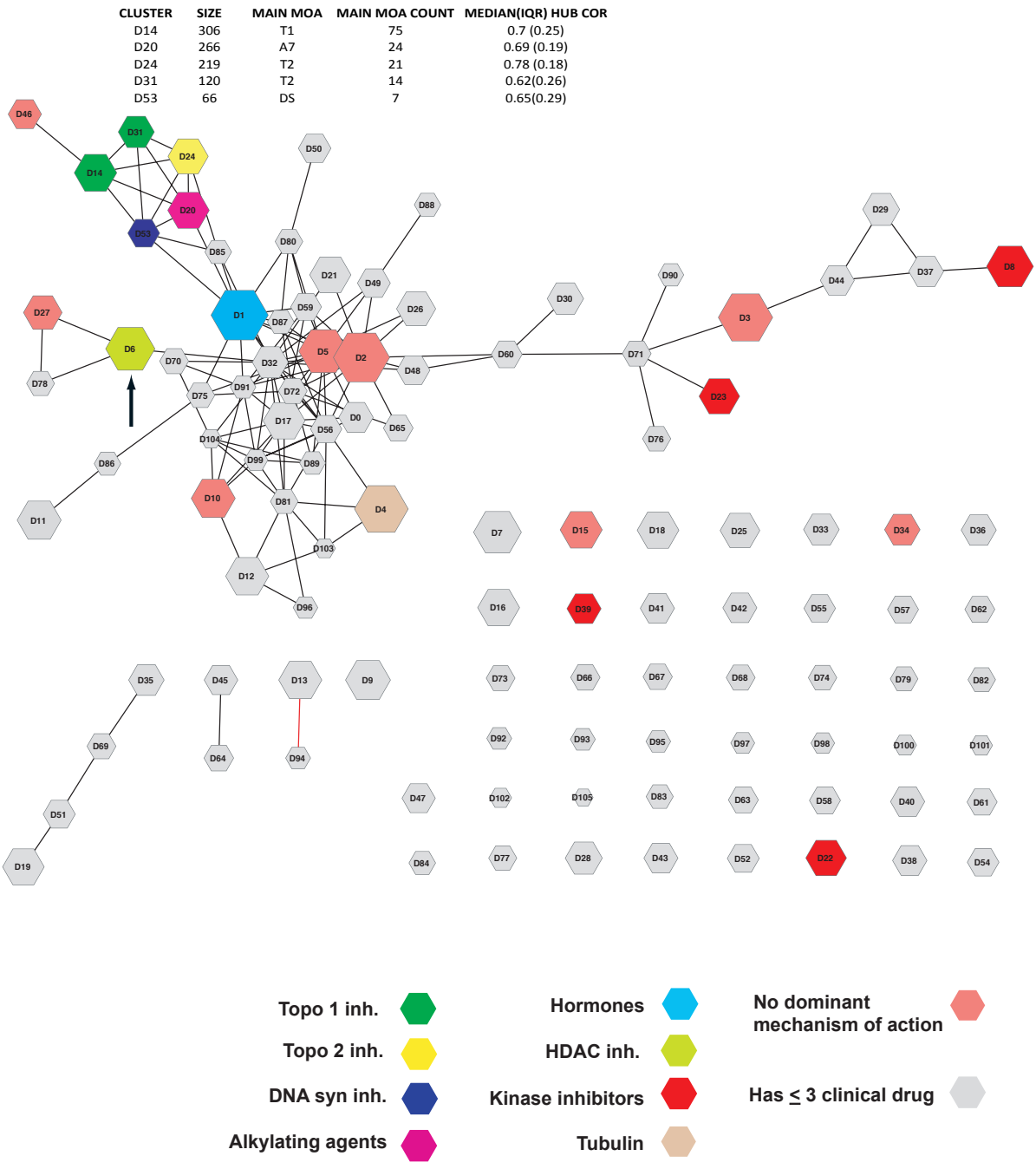


Figure 5.5: Network of 105 drug compound clusters showing clusters containing 3 or more known mechanism of action compounds, together with the dominant compound category. Edges indicate cluster hub-hub correlations greater than 0.6 in magnitude, with positive correlations in black and negative correlations in red.

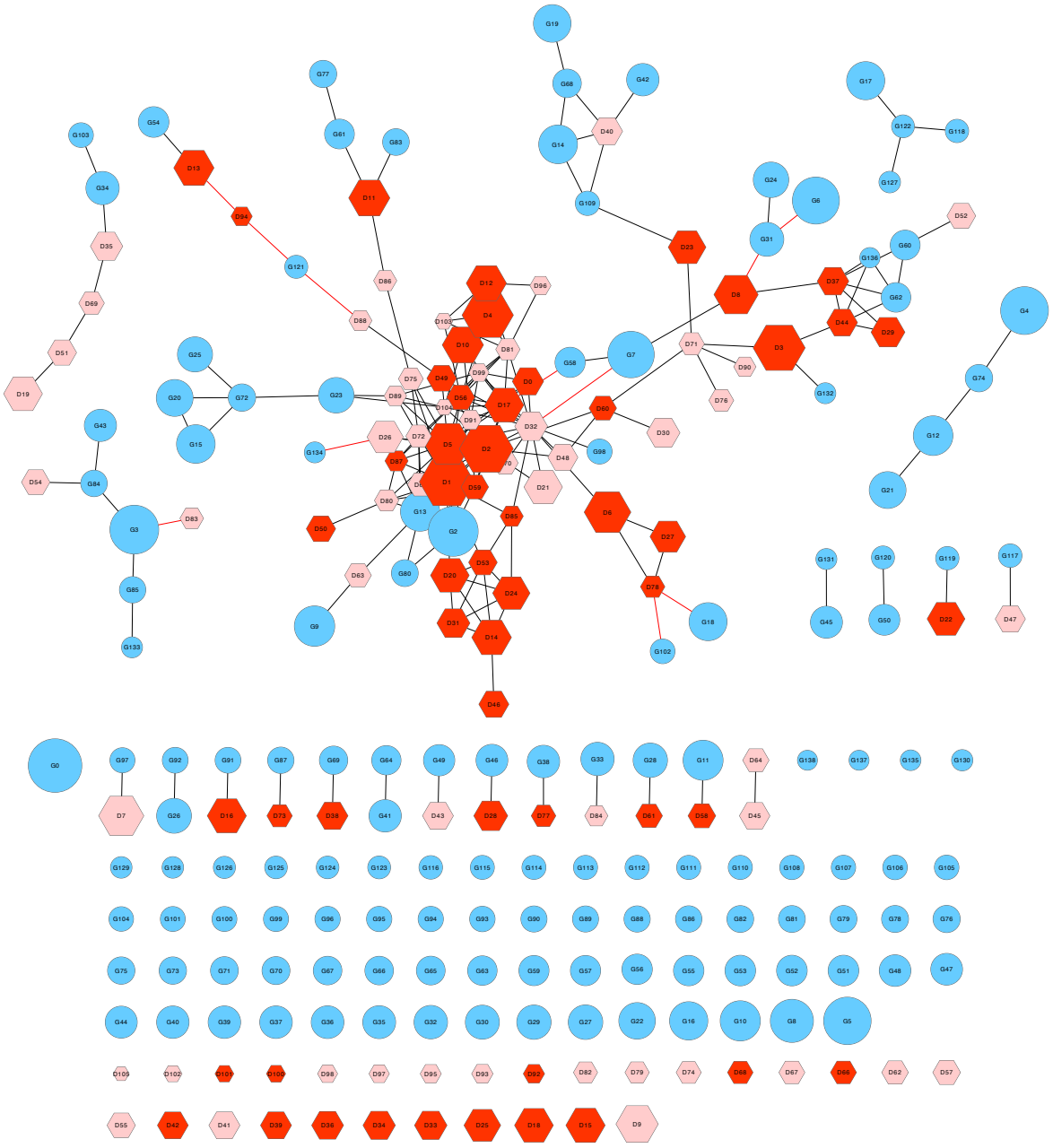


Figure 5.6: Joint network of 105 drug compound clusters and 139 gene co-expression modules, with co-expression modules in blue, drug clusters containing known mechanism of action compounds in dark red, and drug clusters containing only unknown mechanism of action compounds in light red. Edges indicate cluster/module hub-hub correlations greater than 0.6 in magnitude, with positive correlations in black and negative correlations in red.

D10-G2 DNA Damaging Drug Activity - DNA Damage Response Gene Expression Correlations

	266046	271674	363812	RAD51L3	MSH5	TOP1MT
266046	1	0.815	0.811	0.412	0.315	0.428
271674	0.815	1	0.902	0.338	0.375	0.414
363812	0.811	0.902	1	0.374	0.326	0.359
RAD51L3	0.412	0.338	0.374	1	0.357	0.125
MSH5	0.315	0.375	0.326	0.357	1	0.313
TOP1MT	0.428	0.414	0.359	0.125	0.313	1

D53-G126 DNA Damaging Drug Activity - DNA Damage Response Gene Expression Correlations

	63878	105014	145668	287459	606869	FANCA
63878	1	0.674	0.91	0.785	0.604	0.43
105014	0.674	1	0.574	0.607	0.707	0.316
145668	0.91	0.574	1	0.749	0.545	0.365
287459	0.785	0.607	0.749	1	0.708	0.314
606869	0.604	0.707	0.545	0.708	1	0.36
FANCA	0.43	0.316	0.365	0.314	0.36	1

D10-G7 DNA Damaging Drug Activity - DNA Damage Response Gene Expression Correlations

	266046	271674	363812	366140	354646	268242	APLF
266046	1	0.815	0.811	0.483	0.593	0.726	-0.42
271674	0.815	1	0.902	0.389	0.522	0.769	-0.436
363812	0.811	0.902	1	0.361	0.529	0.688	-0.421
366140	0.483	0.389	0.361	1	0.405	0.616	-0.385
354646	0.593	0.522	0.529	0.405	1	0.65	-0.267
268242	0.726	0.769	0.688	0.616	0.65	1	-0.385
APLF	-0.42	-0.436	-0.421	-0.385	-0.267	-0.385	1

D20-G2 DNA Damaging Drug Activity - DNA Damage Response Gene Expression Correlations

	762	3088	6396	8806	9706	34462	132313	296934	RAD51L3	MSH5	DMC1
762	1	0.848	0.782	0.813	0.795	0.856	0.801	0.778	0.416	0.287	0.414
3088	0.848	1	0.927	0.945	0.945	0.978	0.905	0.897	0.352	0.287	0.452
6396	0.782	0.927	1	0.906	0.986	0.935	0.964	0.941	0.341	0.266	0.392
8806	0.813	0.945	0.906	1	0.914	0.947	0.901	0.893	0.28	0.34	0.466
9706	0.795	0.945	0.986	0.914	1	0.957	0.973	0.93	0.371	0.282	0.376
34462	0.856	0.978	0.935	0.947	0.957	1	0.928	0.915	0.352	0.299	0.392
132313	0.801	0.905	0.964	0.901	0.973	0.928	1	0.924	0.367	0.297	0.369
296934	0.778	0.897	0.941	0.893	0.93	0.915	0.924	1	0.331	0.266	0.282
RAD51L3	0.416	0.352	0.341	0.28	0.371	0.352	0.367	0.331	1	0.357	0.238
MSH5	0.287	0.287	0.266	0.34	0.282	0.299	0.297	0.266	0.357	1	0.267
DMC1	0.414	0.452	0.392	0.466	0.376	0.392	0.369	0.282	0.238	0.267	1

Figure 5.7: Pairwise correlations between known DNA damaging drug compounds and DNA damage response genes. Correlations shown in bold are significant at $p < 0.05$, without adjustment for multiple testing.

D23-G6 Kinase Inhibitor Activity - Kinase Gene Expression Correlations

	354462	741078	764042	AXL	DCLK2	NUAK1
354462	1	0.658	0.56	-0.472	-0.357	-0.377
741078	0.658	1	0.952	-0.38	-0.321	-0.416
764042	0.56	0.952	1	-0.273	-0.257	-0.389
AXL	-0.472	-0.38	-0.273	1	0.492	0.603
DCLK2	-0.357	-0.321	-0.257	0.492	1	0.508
NUAK1	-0.377	-0.416	-0.389	0.603	0.508	1

D23-G3 Kinase Inhibitor Activity - Kinase Gene Expression Correlations

	354462	741078	761431	679828	FYN	CDK2	PRKCD	PRKCE	STK10
354462	1	0.658	0.754	0.574	0.338	0.456	0.556	0.51	0.488
741078	0.658	1	0.632	0.629	0.29	0.331	0.637	0.447	0.499
761431	0.754	0.632	1	0.619	0.379	0.499	0.62	0.663	0.495
679828	0.574	0.629	0.619	1	0.37	0.272	0.593	0.426	0.475
FYN	0.338	0.29	0.379	0.37	1	0.466	0.256	0.315	0.517
CDK2	0.456	0.331	0.499	0.272	0.466	1	0.411	0.602	0.331
PRKCD	0.556	0.637	0.62	0.593	0.256	0.411	1	0.527	0.623
PRKCE	0.51	0.447	0.663	0.426	0.315	0.602	0.527	1	0.44
STK10	0.488	0.499	0.495	0.475	0.517	0.331	0.623	0.44	1

Figure 5.8: Pairwise correlations between known kinase inhibitors and kinase targets. Correlations shown in bold are significant at $p < 0.05$, without adjustment for multiple comparisons.

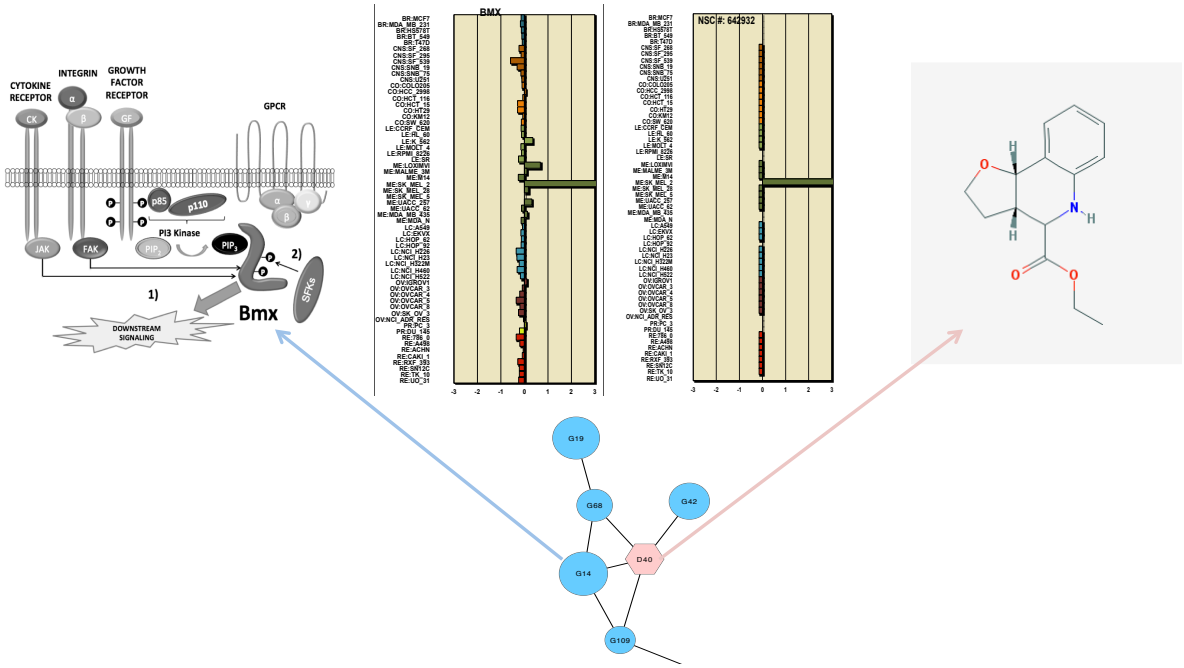


Figure 5.9: Pattern comparison of Bmx hub gene expression profile and NSC642932 hub compound chemoactivity profile. Bmx pathway figure adapted from [44].

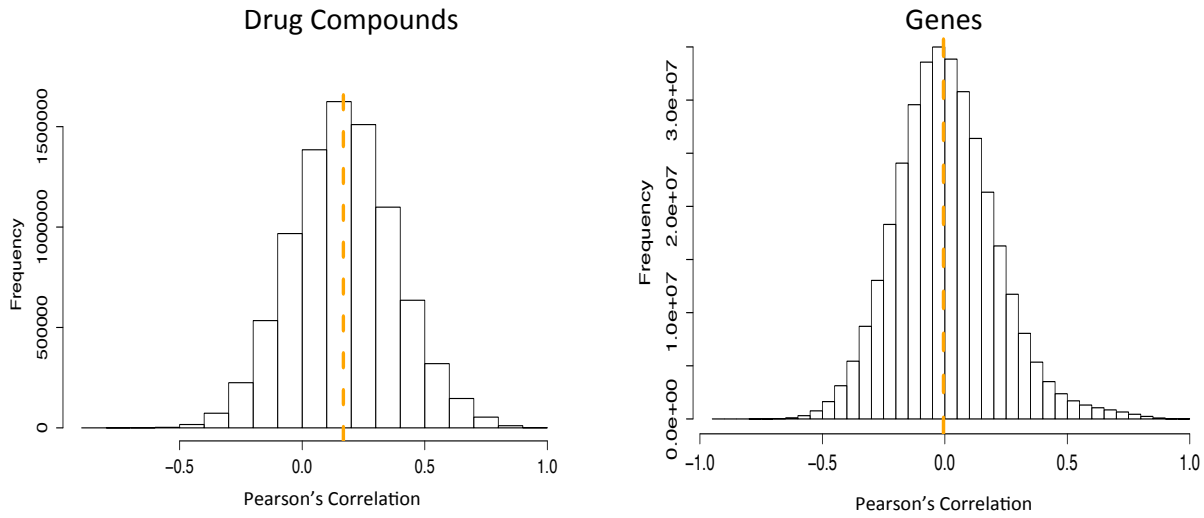


Figure 5.10: Comparison of intra-data class pairwise correlation distributions: drug compound activity profiles versus gene expression profiles.

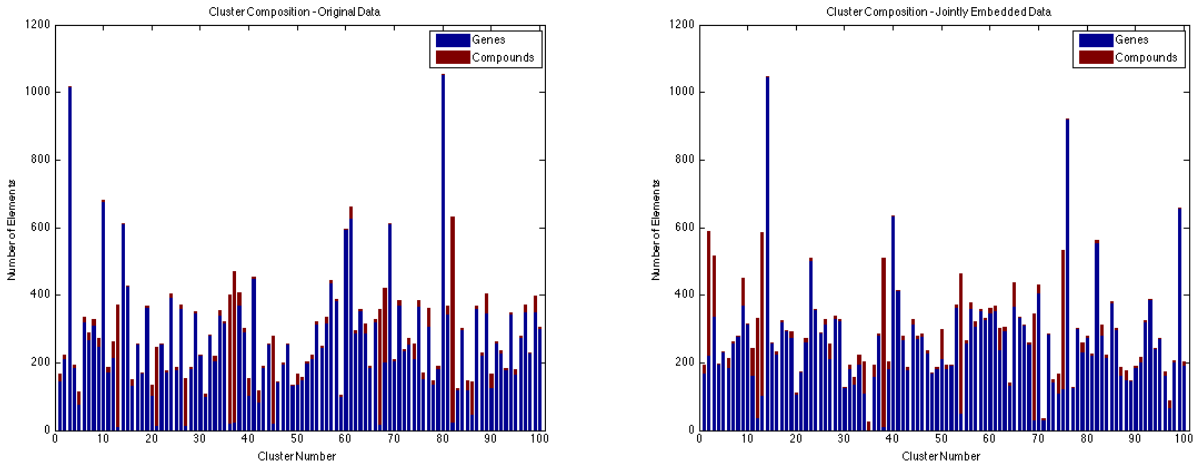


Figure 5.11: Comparison of k-means cluster composition: original drug activity and gene expression profiles (left) versus jointly embedded data (right).

Bibliography

[1] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25, 2000.

[2] C. Bartenhagen, H. Klein, C. Ruckert, X. Jiang, and M. Dugas. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics*, 11(1):567, 2010.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14:585–591, 2001.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[5] M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *J. Comput. Syst. Sci*, 74(8):1289–1308, 2008.

[6] J. Benedetto, W. Czaja, J. Dobrosotskaya, T. Doster, K. Duke, and D. Gillis. Integration of heterogeneous data for classification in hyperspectral satellite imagery. In *SPIE Defense, Security, and Sensing*, pages 839027–839027. International Society for Optics and Photonics, 2012.

[7] J. Benedetto, W. Czaja, J. Dobrosotskaya, T. Doster, K. Duke, and D. Gillis. Semi-supervised learning of heterogeneous data in remote sensing imagery. In *SPIE Defense, Security, and Sensing*, pages 840104–840104. International Society for Optics and Photonics, 2012.

[8] R.F. Bonner, M. Emmert-Buck, K. Cole, T. Pohida, R. Chuaqui, S. Goldstein, L.A. Liotta, et al. Laser capture microdissection: molecular analysis of tissue. *Science (New York, NY)*, 278(5342):1481–1483, 1997.

[9] J.D. Brown, S. Dutta, K. Bharti, R.F. Bonner, P.J. Munson, I.B. Dawid, A.L. Akhtar, I.F. Onojafe, R.P. Alur, J.M. Gross, et al. Expression profiling during ocular development identifies 2 nlz genes with a critical role in optic fissure closure. *Proceedings of the National Academy of Sciences*, 106(5):1462–1467, 2009.

[10] D.J. Cameron, Z. Yang, D. Gibbs, H. Chen, Y. Kaminoh, A. Jorgensen, J. Zeng, L. Luo, E. Brinton, G. Brinton, et al. Htra1 variant confers similar risks to geographic atrophy and neovascular age-related macular degeneration. *Cell Cycle*, 6(9):1122–1125, 2007.

[11] E.J. Candes, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[12] P.G. Casazza and J. Kovačević. Equal-norm tight frames with erasures. *Advances in Computational Mathematics*, 18(2-4):387–430, 2003.

[13] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[14] O. Christensen. *Frames and Bases: An Introductory Course*. Birkhauser Boston, 2008.

[15] F.R.K. Chung. Spectral Graph Theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

[16] B.A. Cohen, R.D. Mitra, J.D. Hughes, and G.M. Church. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genetics*, 26(2):183–186, 2000.

[17] R.R. Coifman and M.J. Hirn. Diffusion maps for changing data. *Applied and Computational Harmonic Analysis*, 2013.

[18] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[19] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426, 2005.

[20] W. Czaja and M. Ehler. Schroedinger eigenmaps for the analysis of biomedical data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1274–1280, 2013.

[21] C. Davis and W.M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[22] D.L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[23] D.L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[24] R.J. Duffin and A.C. Schaeffer. A class of nonharmonic fourier series. *Transactions of the American Mathematical Society*, 72(2):341–366, 1952.

[25] M. Ehler, V.N. Rajapakse, B. Zeeberg, B. Brooks, J. Brown, W. Czaja, and R. Bonner. Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development. *BMC Proceedings*, 5(Suppl 2):S3, 2011.

[26] M. Ehler, V.N. Rajapakse, B. Zeeberg, B.P. Brooks, J.D. Brown, W. Czaja, and R.F. Bonner. Analysis of temporal-spatial co-variation within gene expression microarray data in an organogenesis model. In Mark Borodovsky, Johann Peter Gogarten, Teresa M. Przytycka, and Sanguthevar Rajasekaran, editors, *Bioinformatics Research and Applications, 6th International Symposium, ISBRA 2010, Storrs, CT, USA, May 23-26, 2010. Proceedings*, volume 6053 of *Lecture Notes in Computer Science*, pages 38–49. Springer, 2010.

[27] J.C. Flake. *The Multiplicative Zak Transform, Dimension Reduction, and Wavelet Analysis of LIDAR Data*. PhD thesis, University of Maryland, College Park, Maryland, 2010.

[28] P. Fraser and W. Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413–417, 2007.

[29] Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov. Manifold learning: The price of normalization. *The Journal of Machine Learning Research*, 9:1909–1939, 2008.

[30] S.R. Goldstein, P.G. McQueen, and R.F. Bonner. Thermal modeling of laser capture microdissection. *Applied Optics*, 37(31):7378–7391, 1998.

[31] V.K. Goyal, J. Kovačević, and J.A. Kelner. Quantized frame expansions wit erasures. *Applied and Computational Harmonic Analysis*, 10(3):203–233, 2001.

[32] A. Halevy. *Extensions of Laplacian Eigenmaps for Manifold Learning*. PhD thesis, University of Maryland, College Park, Maryland, 2011.

[33] J. Ham, D.D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 47. ACM, 2004.

[34] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[35] M.A. Harris, J.I. Deegan, J. Lomax, M. Ashburner, S. Tweedie, S. Carbon, S. Lewis, C. Mungall, J. Day-Richter, K. Eilbeck, et al. The gene ontology project in 2008. *Nucleic Acids Research*, 36:D440–D444, 2008.

[36] T.J. Hastie, R.J. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[37] T.J. Hestilow and Y. Huang. Clustering of gene expression data based on shape similarity. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009:3, 2009.

[38] M.J. Hirn. *Enumeration of Harmonic Frames and Frame Based Dimension Reduction*. PhD thesis, University of Maryland, 2009.

[39] T. Holopainen, V. López-Alpuche, W. Zheng, R. Heljasvaara, D. Jones, Y. He, D. Tvorogov, G. D'Amico, Z. Wiener, L.C. Andersson, et al. Deletion of the endothelial bmx tyrosine kinase decreases tumor angiogenesis and growth. *Cancer Research*, 72(14):3512–3521, 2012.

[40] A.L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

[41] S. Horvath and J. Dong. Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, 4(8):e1000117, 2008.

[42] H. Hotelling. Analysis of a complex of statistical variables into principal components. *The Journal of Educational Psychology*, pages 498–520, 1933.

[43] B. Hunter and T. Strohmer. Performance analysis of spectral clustering on compressed, incomplete and inaccurate measurements. *arXiv preprint arXiv:1011.0997*, 2010.

[44] J.S. Jarboe, S. Dutta, S.E. Velu, C.D. Willey, et al. Mini-review: Bmx kinase inhibitors for cancer therapy. *Recent Patents on Anti-Cancer Drug Discovery*, 2012.

[45] K. Karhunen. *Zur spektraltheorie stochastischer prozesse*, volume 34. Suomalainen tiedeakatemia, 1946.

[46] S. T. Kosak and M. Groudine. Form follows function: The genomic organization of cellular differentiation. *Genes & Development*, 18(12):1371–1384, 2004.

[47] J. Kovacevic and A. Chebira. Life Beyond Bases: The Advent of Frames (Part I). *Signal Processing Magazine, IEEE*, 24(4):86–104, 2007.

[48] J. Kovacevic and A. Chebira. Life Beyond Bases: The Advent of Frames (Part II). *Signal Processing Magazine, IEEE*, 24(5):115–125, 2007.

[49] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[50] P. Langfelder, B. Zhang, and S. Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 2008.

[51] P.D. Lax. *Linear Algebra and Its Applications*. Number v. 10 in Pure and Applied Mathematics. Wiley, 2007.

[52] G. Lee, C. Rodriguez, and M. Madabhushi. Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3):368–384, 2008.

[53] J.M. Lee and E.L.L. Sonnhammer. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Research*, 13(5):875–882, 2003.

[54] K. Lee, S. Nam, E. Cho, I. Seong, J. Limb, S. Lee, and J. Kim. Identification of direct regulatory targets of the transcription factor sox10 based on function and conservation. *BMC Genomics*, 9(1):408, 2008.

[55] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[56] H. Liu, P. D'Andrade, S. Fulmer-Smentek, P. Lorenzi, K.W. Kohn, J.N. Weinstein, Y. Pommier, and W.C. Reinhold. mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Molecular Cancer Therapeutics*, 9(5):1080, 2010.

[57] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[58] M. Loève. Fonctions aléatoires de second ordre. *CR Acad. Sci. Paris*, 220:380, 1945.

[59] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[60] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.

[61] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[62] I. Rajapakse, M.D. Perlman, D. Scalzo, C. Kooperberg, M. Groudine, and S.T. Kosak. The emergence of lineage-specific chromosomal topologies from coordinate gene regulation. *Proceedings of the National Academy of Sciences*, 106(16):6679–6684, 2009.

[63] V.N. Rajapakse, W. Czaja, Y.G. Pommier, W.C. Reinhold, and S. Varma. Predicting expression-related features of chromosomal domain organization with network-structured analysis of gene expression and chromosomal location. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 226–233. ACM, 2012.

[64] S. Reichman, R.K.R. Kalathur, S. Lambard, N. Aït-Ali, Y. Yang, A. Lardenois, R. Ripp, O. Poch, D.J. Zack, J.A. Sahel, et al. The homeobox gene chx10/vsx2 regulates rdcvf promoter activity in the inner retina. *Human Molecular Genetics*, 19(2):250–261, 2010.

[65] W.C. Reinhold, J.L. Mergny, H. Liu, M. Ryan, T.D. Pfister, R. Kinders, R. Parchment, J. Doroshow, J.N. Weinstein, and Y. Pommier. Exon array analyses across the NCI-60 reveal potential regulation of TOP1 by transcription pausing at guanosine quartets in the first intron. *Cancer Research*, 70(6):2191, 2010.

[66] W.C. Reinhold, M. Sunshine, H. Liu, S. Varma, K.W. Kohn, J. Morris, J. Doroshow, and Y. Pommier. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Research*, 72(14):3499–3511, 2012.

[67] S. Rosenberg. *The Laplacian on a Riemannian manifold: an introduction to analysis on manifolds*, volume 31. Cambridge University Press, 1997.

[68] P.J. Rousseeuw and L. Kaufman. Finding groups in data: An introduction to cluster analysis. *John, John Wiley & Sons*, 1990.

[69] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[70] B. Schölkopf, A. Smola, and K.R. Müller. Kernel principal component analysis. *Artificial Neural Networks ICANN'97*, pages 583–588, 1997.

[71] U.T. Shankavaram, W.C. Reinhold, S. Nishizuka, S. Major, D. Morita, K.K. Chary, M.A. Reimers, U. Scherf, A. Kahn, D. Dolginow, et al. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, 6(3):820, 2007.

[72] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[73] R.H. Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews*, 6(10):813–823, 2006.

[74] C.L. Sigulinsky, E.S. Green, A.M. Clark, and E.M. Levine. Vsx2/chx10 ensures the correct timing and magnitude of hedgehog signaling in the mouse retina. *Developmental Biology*, 317(2):560–575, 2008.

[75] G. Strang. *Linear Algebra and its Applications*. Thomson Brooks/Cole Cengage Learning, 2006.

[76] A. Sturn, J. Quackenbush, and Z. Trajanoski. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–208, 2002.

[77] C.A. Suarez-Quian, S.R. Goldstein, and R.F. Bonner. Laser capture microdissection: a new tool for the study of spermatogenesis. *Journal of Andrology*, 21(5):601, 2000.

[78] J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[79] J.A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051, 2006.

[80] E. Van Den Berg and M.P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[81] Bert Vogelstein and Kenneth W Kinzler. Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799, 2004.

[82] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[83] J.N. Weinstein, T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace Jr, K.W. Kohn, T. Fojo, S.E. Bates, L.V. Rubinstein, N.L. Anderson, et al. An information-intensive approach to the molecular pharmacology of cancer. *Science*, 275(5298):343–349, 1997.

[84] D. P. Widemann. *Dimensionality Reduction for Hyperspectral Data*. PhD thesis, University of Maryland, 2008.

[85] Y.H. Woo, M. Walker, and G.A. Churchill. Coordinated expression domains in mammalian genomes. *PloS One*, 5(8):e12158, 2010.

[86] Z. Yang, N.J. Camp, H. Sun, Z. Tong, D. Gibbs, D.J. Cameron, H. Chen, Y. Zhao, E. Pearson, X. Li, et al. A variant of the htra1 gene increases susceptibility to age-related macular degeneration. *Science*, 314(5801):992–993, 2006.

[87] Y. Yi, J. Mirosevich, Y. Shyr, R. Matusik, et al. Coupled analysis of gene expression and chromosomal location. *Genomics*, 85(3):401, 2005.

[88] B. Zeeberg, H. Liu, A. Kahn, M. Ehler, V. N. Rajapakse, R. Bonner, J. Brown, B. Brooks, V. Larionov, W. Reinhold, J. Weinstein, and Y. Pommier. RedundancyMiner: De-replication of redundant GO categories in microarray and proteomics analysis. *BMC Bioinformatics*, 12(1):52, 2011.

[89] B. Zeeberg, H. Qin, S. Narasimhan, M. Sunshine, H. Cao, D. Kane, M. Reimers, R. Stephens, D. Bryant, S. Burt, et al. High-Throughput GoMiner, an'industrial-strength'integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*, 6(1):168, 2005.

[90] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, 2003.