Assessing a Nonlinear Dimensionality Reduction-Based Approach to Biological Network Reconstruction.

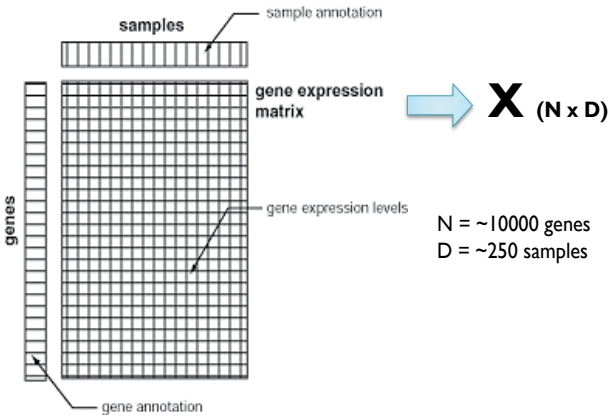
Vinodh N. Rajapakse – vinodh@math.umd.edu
PhD Advisor: Professor Wojciech Czaja – wojtek@math.umd.edu

Project Background and Aims

Deeper understanding of biological processes requires a systems perspective, integrating knowledge of both component elements (genes, proteins, metabolites, etc.), as well as their broader interactions. Graph theory provides a rich language for describing and analyzing these sorts of complex, networked systems. But, reliable inference of biological network structure from noisy, high-dimensional, and still fundamentally limited data remains a challenging problem. A promising avenue involves integrating nonlinear dimensionality reduction (DR) approaches with network analysis and reconstruction algorithms. Beyond basic dimensionality reduction, these approaches can potentially accentuate latent structure and organize data in ways that allow integration of diverse information sources, including prior knowledge. Software tools for performing elements of this analysis exist, but they are distributed over many packages, which often do not interact easily together.

This project will build a basic, integrated data analysis pipeline for deriving gene association network models from gene expression data. Dimensionality reduction techniques will be applied to map the input data in ways that aim to capture intrinsic structure. After this step, standard network reconstruction and analysis techniques will be applied. Algorithm implementations will be individually validated using well-characterized data sets and established software. Following this foundational work, the impact of dimensionality reduction on the overall network reconstruction will be systematically assessed using additional validation and testing data sets. In particular, the initial aim will be to implement, validate, and incorporate a representative nonlinear dimensionality reduction technique – Laplacian Eigenmaps. Network reconstructions derived using data processed by this technique will be compared with ones derived using a leading method, with ones derived directly from the original data, as well as with ones derived from data processed using a standard linear dimensionality reduction approach – Principal Component Analysis. Pending successful completion of this work, some extensions of the basic analysis pipeline are possible. These notably include an approach to enhance handling of very large data sets, as well as consideration of an additional nonlinear dimensionality reduction technique – Diffusion Maps. Further details are provided in subsequent sections of this proposal.

Detailed Approach



The starting point for this analysis is an N x D gene expression data matrix, which we will denote by **X**. The entry $(X)_{ij}$ in **X** records the expression (relative abundance) of the i^{th} gene (mRNA product) in the j^{th} sample. In what follows, we will focus on gene expression profiles across samples, which are naturally organized along the rows of **X**. The expression profile of the i^{th} gene will be designated by $\mathbf{x_i}$, and will be regarded as a point in a D-dimensional input data space. From this input data, network reconstruction will proceed in three broad steps:

(1) Starting from the N x D gene expression data matrix **X**, derive an N x d matrix **Y**, (d < D) using Laplacian Eigenmaps (or another dimensionality reduction technique).
(2) Construct an N x N matrix **W*** capturing pairwise Euclidean distances between row vectors of **Y** (which can be regarded as 'reduced gene profiles').
(3) Apply a statistical significance-based threshold to the elements of **W*** to obtain a network (adjacency matrix) representation.

Laplacian Eigenmaps Background

The motivation for integrating nonlinear dimensionality reduction with network reconstruction and analysis derives from the thought that gene expression data sets might be relatively constrained to lower dimensional manifolds within the high-dimensional spaces in which they reside. This notion is not implausible given the highly ordered structure of gene regulatory networks and their associated processes. Laplacian Eigenmaps (LE) aims to map input space points to a lower dimensional space in such a way that local relationships are preserved (2). These mapped space distances might approximate distances along a putative data manifold in the original space. Some initial studies have suggested that their application might support more biologically specific clustering of gene expression profiles (3). These results motivate the more thorough consideration of manifold learning-based nonlinear dimensionality reduction in support of detailed network structure recovery.

Operationally, Laplacian Eigenmaps transforms an input (N x D) data matrix (data element vectors organized along rows) to an output (N x d) matrix (d < D). There are three steps, though the first two can be combined computationally.

(1) Model Data Point Relationships: Build a graph G with nodes *i* and *j* connected if $\mathbf{x_i}$ is one of the k nearest neighbors of $\mathbf{x_j}$ or vice versa. (Euclidean distances will be used at least initially; alternatives are possible) k is a local structure resolution parameter.
(2) Construct Weight Matrix: Form a diffusion weight matrix **W**, with entry $(W)_{ij} = \exp\{ - || \mathbf{x_i} - \mathbf{x_j} ||^2 \}$ if i and j are connected; 0 otherwise.
(3) Solve the Minimization Problem:

$$\min_{(Y^T DY=I)} \frac{1}{2}\sum_{i,j} \| y_i - y_j \|^2 W_{i,j} = \min_{(Y^T DY=I)} \mathrm{trace}(Y^T LY),$$

In the above formulation, **D** is a diagonal 'connectivity matrix', whose entries record the sum of the edge weights for each data point-derived node (as recorded along the rows or columns of **W**). $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. *The key idea with the minimization is to force points that are close together in the original data space (e.g., that have larger associated edge weights) to lie close together in the mapped data space, while more loosely related points can be relatively pushed apart, drawing out (potentially complex, nonlinear) structure in the data.* The specified constraint serves to scale the output and eliminate trivial solutions. Some basic results from linear algebra show that the optimal mapping can be obtained by solving the generalized eigenvalue problem $\mathbf{Lx} = \lambda\mathbf{Dx}$ under the above constraint. In particular, the coordinates of the mapped data point $\mathbf{y}_i$ in the space of reduced dimension *d* can be extracted from the i^{th} coordinates of the *d* eigenvectors with the smallest nonzero eigenvalues (2).

The Laplacian Eigenmaps method has two major strengths. First, the data-organizing map can be obtained by solving a standard and reasonably tractable computational problem. In addition, the approach comes with deeper theoretical assurances of optimal manifold recovery, in the limit of sufficient data. In particular, the graph-based Laplacian matrix L can be seen as a discrete analogue of the Laplacian-Beltrami operator on the underlying manifold. The eigenmaps of the latter operator can be shown to 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 (2).

Derivation of Network Structure

Given an N x N matrix **W*** capturing pairwise Euclidean distances between mapped gene expression profiles *in the reduced d-dimensional space*, and a target p-value α, a distance threshold for deriving a network adjacency matrix from **W*** will be obtained by applying the following steps.

- Run random data (scaled comparably to input data) through the described dimensionality reduction process.
- Rank pairwise distances between the derived points in the mapped data space.
- Select the distance threshold that excludes the upper $(\mathbf{1} - \alpha)$ fraction of observed distance values.

Additional Details and Possible Extensions

Laplacian Eigenmaps parameter selection: Two parameters must be selected in the course of constructing the data map using Laplacian Eigenmaps – the target reduced dimension d, and the neighborhood resolution parameter k. While in some instances, it may be possible to select the target dimension based on knowledge of the biological process under investigation, for this project an established data-driven maximum likelihood technique will be used to estimate the intrinsic data dimensionality (8). Past work with gene expression data sets of the size considered in this project have yielded good results with neighborhood resolution parameter settings in range of k = 10 to k = 20, which are along the lines of the values applied to example data sets in the original Laplacian Eigenmaps paper (2). For the first phase of the project, values in this range will be used and assessed. Following successful validation and testing of the core algorithm and basic framework, more refined approaches may be implemented and assessed (7).

Computational Resource Management: While the test data sets in this project will contain data associated with around N = 10 to 20 thousand genes, the aim is to implement a basic library that can scale to anticipated larger research data sets. As such, care will be taken to conserve memory resources, by e.g., avoiding explicit construction of large matrices such as **W*** (N x N) described above. In the latter instance, once a suitable distance threshold is computed, only the distances weights for the corresponding neighbors of a node would need to be recorded (in a suitable sparse matrix data structure). Relatively basic steps like these, to utilize memory efficiently, will be considered and implemented in the early stages of the project. More complex areas, should they arise, will be pushed to the second phase of the project, after basic validation and testing. Some basic test suites will be constructed in these cases, so that modifications can be efficiently made and assessed with care to preserve correctness.

Possible Numerical Challenges: At least one reference has suggested that sparse spectral dimensionality reduction techniques like Laplacian Eigenmaps may face eigenvalue problems that can potentially challenge even state of the art eigensolvers, due to the large range between the smallest and largest eigenvalues (10). Since techniques like Laplacian Eigenmaps rely on selection of eigenvectors associated with the d smallest (nonzero) eigenvalues, instances where these cannot be accurately identified could potentially result in suboptimal data mappings. Initial work in this project will apply (a suitable C implementation of) the default eigensolver used in MATLAB, since solid results have been obtained using the latter package. The developed Laplacian Eigenmaps implementation will be parameterized with respect to the eigensolver routine, so that alternative choices can be readily explored as appropriate, if not for this project, in later research work.

Possible Extensions: Pending successful implementation, validation, and testing of the basic Laplacian Eigenmaps-based network reconstruction approach outlined above, some extensions are possible in the second term phase of the project. In particular, approximate nearest neighbor selection algorithms could be incorporated and assessed. These could facilitate handling of much larger data sets by expediting the construction of the initial nearest neighbor-restricted weight matrix utilized by the Laplacian Eigenmaps method (and several related nonlinear dimensionality reduction methods). One possible approach with an available library is described in the reference (4). In addition, an additional dimensionality reduction technique could be implemented and assessed in context of network reconstruction. The Diffusion Maps technique, which can also be reduced to an eigenvalue problem, would be a strong candidate (5).

Implementation Notes

The overall aim will be to build a basic network reconstruction tool that can run on standard desktop hardware, for the data set sizes described. As noted above, care will be taken to allow scalability to larger data sets, through appropriate selection of data structures and algorithms. An additional aim is to utilize free and open source platforms that will allow the code to run on typically encountered systems. In particular, the Python language and platform will be used for its broader libraries, as well as for basic algorithm prototyping. The latter efforts will utilize the SciPy and matplotlib libraries, which provide MATLAB-like numerical computing and visualization capabilities. The Laplacian Eigenmaps technique and other computationally intensive elements of the project will be implemented in C/C++. Libraries will be used for basic linear algebra (to be selected in the first few weeks of the project) and network analysis and visualization (iGraph).

Validation and Testing

Validation of the software will proceed in two stages. First, the Laplacian Eigenmaps algorithm implementation will be tested against an established implementation provided by the MATLAB-based Dimensionality Reduction Toolbox (11). The latter library has been extensively used by our group and others. The aim will be to replicate the published performance of the DR Toolbox Laplacian Eigenmaps implementation over 4 synthetic data sets representing relatively simple nonlinear manifolds in three-dimensional space ('Swiss Roll', 'Broken Swiss Roll', 'Twinpeaks', and 'Helix'). These sorts of data sets are commonly used in basic assessment of nonlinear dimensionality reduction techniques. The authors specify two measures – 'trustworthiness' and 'continuity' – for assessing the quality of a lower dimensional data representation. For a given nearest neighbor (neighborhood resolution) parameter setting k, trustworthiness is defined by:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} (r(i,j) - k)\,,$$

where r(i,j) represents the node *i* neighborhood group rank of the data point *j*, according to the pairwise distances from the data point *i* in the high dimensional space, and $U_i^{(k)}$ is the set of points that are among the k nearest neighbors of point *i* in the low dimensional space *but not in the high dimensional space*. The trustworthiness measure ranges from 0 to 1, and *in essence, it aims to penalize mappings that often group points in the low dimensional space that were not close together in the original input space*. The continuity measure is analogously defined to penalize

instances in which points that were close neighbors in the original input space are placed far apart in the mapped space (6, 10). *The basic validation of the C/C++ Laplacian Eigenmaps algorithm implementation will aim to replicate the published continuity and trustworthiness measures for the described artificial data sets (10).* The latter are publicly available, together with applied algorithm parameter settings. Since the MATLAB-based DR Toolbox has been successfully used in several published studies by our group and others, basic alignment of results with these benchmark data sets will give confidence in the new implementation. Analogous results are published for Diffusion Maps and several other dimensionality reduction routines, allowing their implementations to be similarly validated (10).

After basic validation of the Laplacian Eigenmaps algorithm implementation as described, derived network reconstructions will be validated by comparison with published results using the leading ARACNE algorithm, which has been featured in several strong publications (1, 9). In particular, two levels of integrated system testing will be performed. First, the reconstruction results obtained using data derived from simulation of a realistic artificial gene network will be assessed (9). The results here can be precisely and objectively scored according to recovery of known edges. Second, results over a well-studied biological data set focusing on the cancer gene *myc* will be considered. Here, proper resolution of a coherent network module around *myc* will be assessed, along with its elements. Since many genes belonging to the latter group are known through extensive past experimental work, the results can be once again scored with relative objectivity (for what is still a real-world experimental data set). In particular, results can be compared with those published for the leading ARACNE network reconstruction algorithm (1, 9).

In addition to comparisons with the above leading network reconstruction technique, we will build and compare networks derived using a 'vanilla version' of the reconstruction workflow (which forgoes dimensionality reduction), as well as one which incorporates dimensionality reduction based on the more basic, linear Principal Components Analysis technique. The overall results should provide an initial assessment of whether nonlinear dimensionality reduction approaches have particular value in accurately recovering complex biological network structures.

Project Schedule and Milestones

- Phase I: Laplacian Eigenmaps + Basic Network Reconstruction
  - Target Date: early December 2010
  - Milestones:
    - Implementation and focused validation of Laplacian Eigenmaps.
    - Implementation of statistical significance-based distance matrix thresholding for network reconstruction
- Phase II: Integrated Testing of Network Reconstruction + Possible Extensions
  - Target Date: end of March 2011
  - Milestones:
    - Integrated testing of network reconstruction, method comparisons
    - Possible: Approximate Nearest Neighbor Selection Algorithms
    - Possible: Enhanced tuning of Laplacian Eigenmaps algorithm parameters
    - Possible: Implementation and integrated testing of alternative dimensionality reduction techniques (Diffusion Maps)

Deliverables

- Technical Report outlining:
  - Problem and general approach
  - Algorithm and system implementation details of particular interest
  - Validation and testing results, including detailed comparative assessment of nonlinear dimensionality reduction in gene network reconstruction
- Source Code – together with data sets and scripts for reproducing results presented in technical report.

REFERENCES

1. K. Basso, A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, A. Califano *Reverse engineering of regulatory networks in human B cells* Nature Genetics 37, 382 - 390 (2005)
2. M. Belkin and P. Niyogi, *Laplacian Eigenmaps for dimensionality reduction and data representation, Neural Computation.* 15 (2004), No. 6, 1373-1396
3. M. Ehler, V. Rajapakse, B. Zeeberg, B. Brooks, J. Brown, W. Czaja, and R. F. Bonner, *Analysis of temporal-spatial co-variation within gene expression microarray data in an organogenesis model.* 6th International Symposium on Bioinformatics Research and Applications (ISBRA'10), Lecture Notes in Bioinformatics, Springer Verlag, 2010
4. ***J. Chen, H. Fang, Y. Saad*** *Fast Approximate kNN Graph Construction for High Dimensional Data via Recursive Lanczos Bisection* Journal of Machine Learning Research 10 (2009) 1989-2012
5. RR Coifman, S Lafon, A Lee, M Maggioni, B Nadler, FJ Warner, and SW Zucker.*Geometric diffusions as a tool for harmonic analysis and structure definition of data. part i: Diffusion maps* , Proc. of Nat. Acad. Sci., 102:7426--7431, May 2005.
6. Kaskie S, Nikkila J, Oja M, Venna J, Törönen P, Castrén E. *Trustworthiness and metrics in visualizing similarity of gene expression.* BMC Bioinformatics. 2003 Oct 13;4:48.
7. R. Karbauskaite, O. Kurasova, G. Dzemyda *Selection of the number of neighbors of each data point for the locally linear embedding algorithm.* Information Technology and Control, 2007, Vol. 36, No. 4
8. E. Levina and P.J. Bickel. *Maximum likelihood estimation of intrinsic dimension.* In Advances in Neural Information Processing Systems, volume 17, Cambridge, MA, USA, 2004. The MIT Press.
9. A. A. Margolin, I Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R Dalla Favera, A. Califano, *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics. 2006 Mar 20;7 Suppl 1:S7.
10. van der Maaten, Postma, van den Herik, *Dimensionality Reduction: A Comparative Review.* Tilburg Centre for Creative Computing Technical Report 2009-005
11. http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
12. http://igraph.sourceforge.net/