gMap: extracting and interactively visualizing nonlinear relationships of genes from expression

Chaolin Zhang[1], Yanda Li[1], Xuegong Zhang[1,*]

Keywords: gMap, microarray, interactive visualization, nonlinear projection, Isomap

1 Introduction.

DNA micro-array technologies have made it possible to monitor the expression of thousands of genes across a set of conditions or cases simultaneously, enabling us to study the functional or regulatory relationship among genes. Assuming that genes with similar expression profiles have similar functions or are in the same regulatory pathway, various methods of gene clustering and dimension reduction by projection are commonly used to extract and represent the underlying knowledge. Two critical problems of these methods are measuring the relations between genes and visualizing these discovered relations in a way that is easy for biologists' analysis.

Due to the high-complexity of gene systems, distance metrics such as correlation and Euclidean distance may not always capture the relationships among genes in the event of time shifts or the different timing rates of biological processes, etc [1].

In these cases, nonlinear manifold might exist in the high dimensional space of gene expression data. By measuring the distance of genes with geodesic distance, the nonlinear manifold can be linearized and preserved more faithfully when the data is projected into the low dimensional space. Zhou, *et al* [3] demonstrated the effectiveness of "shortest path analysis" (which is actually an equivalence of geodesic distance) in transitive functional annotation. In our study, we employed a more systematic framework of applying geodesic distance to extract the relationship among genes. An easy-to-use tool named gMap (*g* means both *geodesic* and *gene*) was developed for this purpose. Another important advantage of gMap is that it includes a number of interactive features to facilitate the scrutiny of biologists upon the expression data.

2 Methods and interactive features.

At the center of gMap is the Isomap algorithm proposed by Tenenbaum, *et al* [2]. Instead of measuring the similarity/dissimilarity of genes with correlation or Euclidean distance (or with any other distance metric) directly, we use their geodesic distance estimated by their shortest path. The authors of [2] proved that when sufficient data (in our context, number of genes) are given, the estimated geodesic distance can be arbitrarily accurate.

After obtaining the geodesic distance matrix of all gene pairs, we project the relationship of genes into the low dimensional embedding. Particularly for micro-array data analysis, gMap can display genes in a 2D or 3D scatter-plot composed of 2 or 3 major components. A number of flexible interactive features are developed based on the scatter-plots, such as capture, query, sort, cluster, etc, which facilitate the interpretation of results. A part of them is list as follows: (1)Capture. To

[1] MOE Key Lab of Bioinformatics / Dept of Automation, Tsinghua University, Beijing 100084, China.

[*] Corresponding author, email: zhangxg@tsinghua.edu.cn. This work is partially supported by NSFC under grant numbers 60275007 and 60234020.

interpret the biological meaning of a component or the local pattern of a set of genes in the embedding, users can select any gene(s). These selected/captured genes are marked out in the scatter-plot and more information of them, such as gene name, descriptions and expression profiles can be given and visualized. (2) Query. When users have a set of interested genes in advance, they can examine/query these genes to investigate their distribution in the embedding and in turn their relationship. (3)Sort. This feature is extremely useful for cell-cycle data. It allows users to re-order all (cell-cycle related) genes according to their phases in the life period.

3 Applications and results.

To demonstrate the utility of gMap, we applied our tool to a simulated dataset and three typical real micro-array datasets. The data sets are: CAKE model (a simulated genetic network), colon cancer, yeast cycle and response of human fibroblasts to serum. References of data source are not listed here due to the limit of space.

We found that gMap can preserve more information in the low dimensional space because of the linearization effect of geodesic distance, which makes it possible to identify weak relationships or patterns in terms of correlation or Euclidean distance measure. Figure 1 shows the comparison of gMap and a previous classical method MDS in the residual variance of projection, which reflects ratios of information preserved after projection. The residual when the order is 2 or 3 by gMap is less than the counterpart by MDS, especially when array number is not so small. Larger residual at the tail is due to the approximation of geodesic distance. We also observe clusters of genes are more compact in the low dimensional embeddings by gMap compared with MDS.

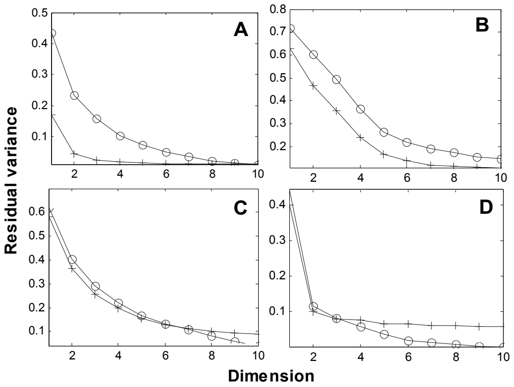

Figure 1: Comparison of gMap and MDS in the residual variance of projection. (+) gMap and (o) MDS. A. CAKE model, B. colon cancer, C. yeast cell cycle, D. response of human fibroblasts to serum

References

[1] Bar-Joseph,Z., Gerber,G., et al. (2003). Continuous Representations of Time Series Gene Expression Data. *J. Comput. Biol.* 3-4, 341-356

[2] Tenenbaum, J. B., Silva,V. D. et al. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*. 290, 2319-2323.

[3] Zhou, X., Kao, M. J. and Wong, W. H. 2002, Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*. 99, 12783-12788