Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Phenotypic-Specific Gene Module Discovery using a Diagnostic Tree and caBIG^{TM} VISDA

Yitan Zhu[1], Zuyi Wang[2,1], Yuanjian Feng[1], Jianhua Xuan[1],
David J. Miller[3], Eric P. Hoffman[2], and Yue Wang[1]

1. Dept. of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA, USA
2. Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC, USA
3. Dept. of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Abstract – **For the critical task of gene module discovery in genomic research, we present a model-based hierarchical data clustering and visualization algorithm, VIsual Statistical Data Analyzer (VISDA), which effectively exploits human-data interaction to improve the clustering outcome. Guided by a diagnostic tree, we apply VISDA to a muscular dystrophy dataset that contains a number of different phenotypic conditions. We then superimpose existing knowledge of gene regulation and gene function (Ingenuity Pathway Analysis) to analyze the clustering results and generate novel hypotheses for further research on muscular dystrophies.**

Keywords – **Gene clustering, hierarchical mixture model, data visualization, gene module, gene regulatory network**

I. Introduction

Microarray technologies and gene expression data provide information that enhances the understanding of functional genomics related to human diseases. One emerging concept is that of the *gene module* comprising a subset of co-regulated genes. To extract information about gene modules from gene expression data, a reasonable approach is to find co-expressed gene clusters, which, putatively, are co-regulated modules. Such a gene module is believed to fulfill some specific biological function, and often exhibits similar patterns across different conditions. Discovering such gene groups is a key to dissecting the gene regulation mechanism in pathway signaling and networking.

Many clustering methods have been recently proposed for module discovery. For example, Tamayo et al. used Self-Organizing Maps (SOM) to find gene clusters with distinct patterns of change in gene expression level for the yeast cell cycle process [1]. For the same task, Sharan and Shamir developed CLICK, a graph-theoretic and statistical algorithm [2]. Biclustering, which performs simultaneous clustering of genes and conditions, is also utilized to find coexpressed gene clusters only under certain conditions [3].

In this paper, we introduce our recently developed data clustering and visualization algorithm, VIsual Statistical Data Analyzer (VISDA) [4]–[6], which characterizes the data with a hierarchical Standard Finite Normal Mixture (SFNM) model and which exploits the human gift for pattern recognition to improve clustering performance (VISDA is a toolkit of caBIG^{TM}). We apply VISDA to a microarray gene expression dataset consisting of 12 different human muscular dystrophies and normal human muscle [7], for the purpose of discovering gene modules and gene regulatory networks that are relevant to the pathogenesis of muscular dystrophies. Specifically, we perform gene clustering under two scenarios: 1) with all the phenotypes as conditions; 2) with specific phenotypes as conditions. Scenario 1 is expected to find gene modules with relatively purer functionality, whereas scenario 2 aims to identify the gene modules specifically involved in certain phenotypes. Since groups of phenotypic conditions are selected according to their similarity, as reflected in a pathologically plausible diagnostic tree, scenario 2 is consistent with the idea of finding homogenous conditions in biclustering. In the rest of this paper, we first discuss the principles behind VISDA and its algorithm implementation. We then introduce the dataset and analyze the clustering results obtained by VISDA, followed by a brief discussion.

II. Algorithm of Visda

Based on a hierarchical SFNM model, VISDA performs top-down divisive clustering as outlined in Fig. 1. At the top level, the whole dataset is split into several coarse clusters that may contain multiple phenotypes; at lower levels, these coarse clusters are further decomposed into finer clusters, until no substructure can be found. This top-down divisive clustering procedure is consistent with the spirit of "divide and conquer". For each cluster at its present level, VISDA focuses on it and projects it onto 2-D spaces using complementary structure-preserving projection methods, which capture different characteristics of the data's structure. To fully take advantage of human intelligence, VISDA allows the user to first select the projection that he/she thinks best reveals the data structure, and then to initialize the sub-cluster centers in this projection space. Through the projection space, VISDA uses Minimum Description Length (MDL) criterion to detect the number of sub-clusters existing in the cluster [8]. Although local data structure may not be detected at the present level, it will become the main data structure captured in sub-clusters at lower levels. Below we provide more detailed explanation for the algorithm.

A. Hierarchical SFNM model and EM algorithm

Let $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be the data points in R^p space, p is the number of dimension. A hierarchical SFNM model uses the following probability density function to describe the relationship between two successive levels in the hierarchy:

1-4244-0033-3/06/$20.00 ©2006 IEEE.

$$\mathrm{f}\left(\mathbf{x}\middle|\boldsymbol{\theta},\boldsymbol{\pi}\right)=\sum_{k=1}^{K_0}\pi_k\sum_{j=1}^{L_k}\pi_{j|k}\,\mathrm{g}\left(\mathbf{x}\middle|\boldsymbol{\theta}_{j|k}\right)\quad\sum_{k=1}^{K_0}\pi_k=1\ \&\ \sum_{j=1}^{L_k}\pi_{j|k}=1\ (1),$$

where K_0 clusters exist in the upper level, L_k sub-clusters exist in the lower level for cluster k, π_k is the mixing proportion for cluster k in the upper level, $\pi_{j|k}$ is the mixing proportion for sub-cluster j within cluster k, g(•) is the Gaussian probability density function, and $\boldsymbol{\theta}_{j|k}$ are the associated parameters.

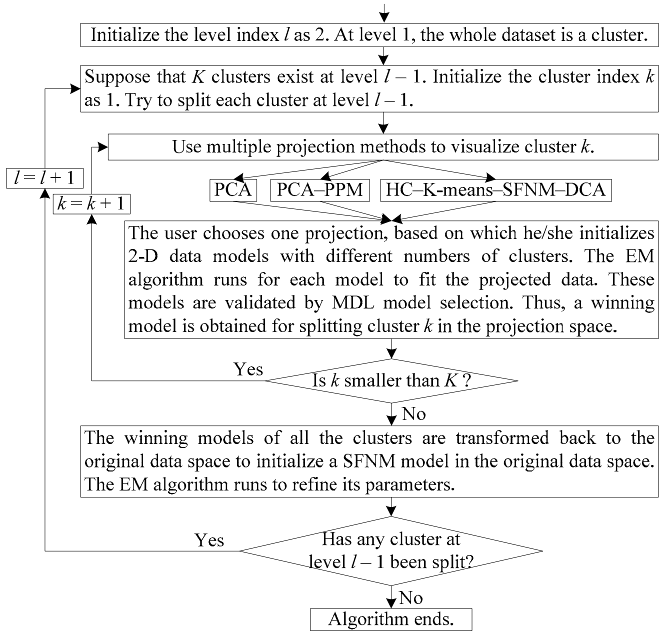
Initialize the level index l as 2. At level 1, the whole dataset is a cluster.
Suppose that K clusters exist at level $l-1$. Initialize the cluster index k as 1. Try to split each cluster at level $l-1$.
Use multiple projection methods to visualize cluster k.
$l = l + 1$
$k = k + 1$
PCA
PCA–PPM
HC–K-means–SFNM–DCA
The user chooses one projection, based on which he/she initializes 2-D data models with different numbers of clusters. The EM algorithm runs for each model to fit the projected data. These models are validated by MDL model selection. Thus, a winning model is obtained for splitting cluster k in the projection space.
Yes
Is k smaller than K ?
No
The winning models of all the clusters are transformed back to the original data space to initialize a SFNM model in the original data space. The EM algorithm runs to refine its parameters.
Yes
Has any cluster at level $l-1$ been split?
No
Algorithm ends.

Fig. 1. The flowchart of VISDA

At each level of the hierarchy, VISDA uses the Expectation Maximization (EM) algorithm to learn the model parameters [9]. The EM algorithm iteratively performs two steps, i.e. the E step and the M step, to monotonically increase the log-likelihood of the data. The E step calculates the conditional expectation of sample labels

$$z_{i(j|k)}=z_{ik}\pi_{j|k}\,\mathrm{g}(\mathbf{x}_i|\boldsymbol{\mu}_{j|k},\mathbf{c}_{j|k})\Big/\sum_{j=1}^{L_k}\pi_{j|k}\,\mathrm{g}(\mathbf{x}_i|\boldsymbol{\mu}_{j|k},\mathbf{c}_{j|k})\qquad(2),$$

where $z_{i(j|k)}$ is the posterior probability of sample $\mathbf{x}_i$ belonging to cluster k and sub-cluster j, z_{ik} is the posterior probability of sample $\mathbf{x}_i$ belonging to cluster k, $\boldsymbol{\mu}_{j|k}$ is the mean of sub-cluster j, and $\mathbf{c}_{j|k}$ is the covariance matrix of sub-cluster j. The M step is

$$\begin{gathered}\pi_{j|k}=\sum_{i=1}^{N}z_{i(j|k)}\Big/\sum_{i=1}^{N}z_{ik}\qquad\boldsymbol{\mu}_{j|k}=\sum_{i=1}^{N}z_{i(j|k)}\mathbf{x}_i\Big/\sum_{i=1}^{N}z_{i(j|k)}\\ \mathbf{c}_{j|k}=\sum_{i=1}^{N}z_{i(j|k)}\left(\mathbf{x}_i-\boldsymbol{\mu}_{j|k}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_{j|k}\right)^T\Big/\sum_{i=1}^{N}z_{i(j|k)}\end{gathered}\qquad(3).$$

This algorithm supports hierarchical cluster decomposition by keeping the sample's probability of belonging to the upper level cluster unchanged and adjusting the conditional probabilities of belonging to the lower level sub-clusters. If the purpose is solely to refine the parameters of an existing SFNM model, we can set K_0 equal to 1. Thus (1) becomes a single level SFNM model. The associated single level EM algorithm is still given by (2) and (3), but with z_{i1} equal to 1. The data model and learning algorithm in the projection space also follows (1), (2), and (3), with the substitution of the projected data, and the mean and covariance matrix of the sub-cluster in the projected space.

B. Complementary Structure-Preserving Projections

The data are projected onto 2-D space by three projection methods, which focus on different characteristics of the data structure. These three projection methods are Principal Component Analysis (PCA), Principal Component Analysis – Projection Pursuit (PCA–PPM) [5], and HC–K-means–SFNM–DCA, where HC refers to Hierarchical Clustering, K-means refers to K-means clustering, and DCA refers to Discriminatory Component Analysis.

PCA finds the two eigenvectors associated with the largest eigenvalues of the cluster's covariance matrix. This projection is optimal for minimizing mean squared reconstruction error. However, minimizing mean squared reconstruction error may not ensure the best preservation of data structure. Since flat distributions or distributions with thick tails usually show some data structure and kurtosis is a good measure of the peakedness of a distribution, the PCA–PPM finds the two eigenvectors on which the projected data distribution has the smallest kurtosis.

The HC–K-means–SFNM–DCA projection takes four steps. 1) Apply HC on the samples that most likely belong to the cluster. The user chooses a distance threshold to cut the samples into sub-clusters. Very small sub-clusters are merged into their nearest larger sub-clusters. 2) Run K-means clustering on the samples using the sub-clusters obtained from HC as initialization. 3) Use the result of K-means clustering to initialize an SFNM model for the cluster, and run the EM algorithm to refine it. 4) DCA projection is done based on the obtained SFNM model. DCA finds the two eigenvectors associated with the largest eigenvalues of Fisher's scatter matrix. To achieve an affine projection, we need to orthogonalize the eigenvectors to get the projection matrix. As we can see, the HC–K-means–SFNM steps obtain an unsupervised partition of the cluster, and DCA allows the partitioned groups to be visualized. If the partition indeed captures the data structure, the DCA projection should show good separability among the projected sub-clusters.

C. Model Selection and Model Transform

The user needs to initialize 2-D SFNM models with different numbers of sub-clusters in the chosen projection space by clicking on the computer screen at the visualized centers of the sub-clusters. Then the EM algorithm runs to refine the models to fit the projected data. The MDL criterion is used to select the best model with the shortest description length. The description length is calculated by

$$\mathrm{L}(\mathbf{Y}) + K_a \log\left(N\right)/2,$$

where **Y** is the projected cluster, L(**Y**) is the log-likelihood of the projected cluster and K_a is the number of free adjustable parameters in the model. To fully take advantage of human intelligence, the user is allowed to override the MDL model selection.

The winning model in the 2-D projection space needs to be transformed back to the original data space to initialize the data model in that space. This transform is achieved by

$$\boldsymbol{\mu}_{j|k} = \mathbf{w}_k \boldsymbol{\mu}_{\mathbf{y},j|k} + \boldsymbol{\mu}_k - \mathbf{w}_k \mathbf{w}_k^T \boldsymbol{\mu}_k \qquad \mathbf{c}_{j|k} = \mathbf{w}_k \mathbf{c}_{\mathbf{y},j|k} \mathbf{w}_k^T,$$

where $\mathbf{w}_k$ is the projection matrix, and $\boldsymbol{\mu}_{\mathbf{y},j|k}$and $\mathbf{c}_{\mathbf{y},j|k}$ are the sub-cluster mean and covariance matrix, respectively, in the projection space.

III. Experimental Results

The muscular dystrophy dataset consists of 13 phenotypes with 125 biopsies and 22215 genes. Table 1 gives a brief summary of the dataset. Fig. 2 shows a pathologically plausible diagnostic tree of phenotypes [7]. The closer two phenotypes are in the tree, the more pathologically or clinically similar they are. Gene clustering was performed on node 1, i.e. under all phenotypic conditions, and on nodes 5 and 7, for which only a subset of the phenotypic conditions was considered. Each gene's expression levels were standardized with respect to the involved phenotypic conditions before clustering. Among the obtained gene clusters, we selected the ones that can differentiate certain phenotypes [1] and used Ingenuity Pathway Analysis (IPA) [10] to assess their biological plausibility, with respect to known information about gene regulatory networks, pathways, and module function.

TABLE I
Muscular Dystrophy Dataset

Phenotype	Sample Number	Description
JDM	25	Juvenile dermatomyositis
FKRP	7	Fukutin related protein deficiency
DMD	10	Duchenne muscular dystrophy, dystrophin deficiency
BMD	5	Becker muscular dystrophy, hypomorphic for dystrophin
Dysferlin	10	Dysferlin deficiency, putative vesicle traffic defect
Calpain III	10	Calpain III deficiency
FSH	14	Fascioscapulohumeral dystrophy
AQM	5	Acute quadriplegic myopathy
HSP	4	Spastin haploinsufficiency, microtubule traffic defect
Lamin A/C	4	Emery dreifuss muscular dystrophy, missense mutations
Emerin	4	Emery dreifuss muscular dystrophy, emerin deficient
ALS	9	Amyotrophic lateral sclerosis
NHM	18	Normal skeletal muscle

We found that the 93rd gene cluster obtained at node 1 is consistently down-expressed in all the phenotypes except for JDM. JDM is a relatively severe childhood autoimmune disorder. It is thought to be associated with viral infections that stimulate muscle destruction by inflammatory cells and ischemic processes in a small subset of the children with the virus. We input the genes most likely belonging to this cluster into IPA. IPA showed that gene regulatory networks involved in key inflammatory pathways were highly ranked by significance. In the top ranked network, shown in Fig. 3, specific proteins that are known to be critical for initiating and perpetuating inflammation and subsequent cell death are seen as key focus genes. STAT1 is an important signaling molecule that responds to interferons and other cytokines. It is highly studied, as suggested by the many nodes emanating from STAT1. Both TNFSF10 (also called TRAIL) and CASP7 (caspase 7) influence cell death via apoptosis. Consistent with this, patients with JDM show extensive cell death and failure of regeneration in their muscle, leading to weakness. This network also points to drugs that would be expected to inhibit this process in JDM patients, which can be tested in mouse models of inflammatory muscle disease.

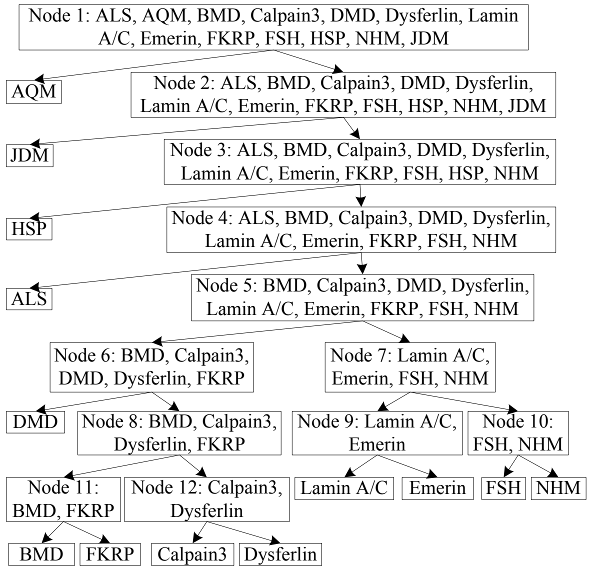
Node 1: ALS, AQM, BMD, Calpain3, DMD, Dysferlin, Lamin A/C, Emerin, FKRP, FSH, HSP, NHM, JDM
AQM
Node 2: ALS, BMD, Calpain3, DMD, Dysferlin, Lamin A/C, Emerin, FKRP, FSH, HSP, NHM, JDM
JDM
Node 3: ALS, BMD, Calpain3, DMD, Dysferlin, Lamin A/C, Emerin, FKRP, FSH, HSP, NHM
HSP
Node 4: ALS, BMD, Calpain3, DMD, Dysferlin, Lamin A/C, Emerin, FKRP, FSH, NHM
ALS
Node 5: BMD, Calpain3, DMD, Dysferlin, Lamin A/C, Emerin, FKRP, FSH, NHM
Node 6: BMD, Calpain3, DMD, Dysferlin, FKRP
Node 7: Lamin A/C, Emerin, FSH, NHM
DMD
Node 8: BMD, Calpain3, Dysferlin, FKRP
Node 9: Lamin A/C, Emerin
Node 10: FSH, NHM
Node 11: BMD, FKRP
Node 12: Calpain3, Dysferlin
Lamin A/C
Emerin
FSH
NHM
BMD
FKRP
Calpain3
Dysferlin

Fig. 2. The pathologically plausible diagnostic tree of the 13 phenotypes.

The 52nd gene cluster obtained at node 5 points to an anabolic IL15 pathway more strongly expressed in FKRP and BMD than in Dysferlin and Calpain3. We hypothesize that the better preservation of muscle histology and function seen in older FKRP and BMD patients, relative to Dysferlin and Calpain3 patients, is in part due to the stronger induction of the anabolic IL15 pathway in the structural membrane dystrophies.

Besides examining the gene regulatory networks and pathways related to the detected modules, we also evaluated individual module's statistical significance in associating with gene function categories, where some interesting results

[1] The differentiability is assessed visually and by the signal to noise ratio.

were also found. For example, the 72nd cluster obtained at node 1, the 55th cluster obtained at node 5, and the 23rd cluster obtained at node 7 have p-values of 4.49E–13, 6.52E–22, and 3.79E–21 associated with belonging to the protein synthesis function category (which is closely related to several muscular dystrophies). These p-values are computed by IPA, based on the hypergeometric distribution, via Fisher's Exact Test for 2×2 contingency tables [10].

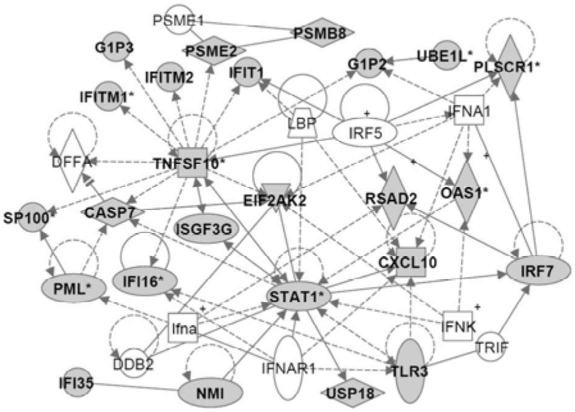

Fig. 3. The top ranked gene regulatory network pointed by the 93rd gene cluster obtained at node 1. Genes with bold names are in the gene cluster. Solid lines indicate direct interactions. Dashed lines indicate indirect interactions. This figure is from IPA system.

IV. Discussion

Unlike automatic clustering methods, for which the user can only adjust input parameters, VISDA provides various human-data interactions during the clustering process, which not only exploits the human gift for pattern recognition, but also is an efficient way to incorporate domain knowledge when used by domain experts. To achieve optimum performance, the user needs to acquire experience in using VISDA on various kinds of data, especially on the dataset of interest. It's also helpful if the user can practice VISDA on labeled datasets (i.e. use the supervised mode).

In the VISDA clustering result, each sample is assigned probabilistic membership in every cluster. Some genes may belong to one cluster with much higher probability than to the other clusters, clearly indicating the gene's category. Other genes may not have a clear tendency for belonging to any cluster. The existence of such genes is consistent with the biological fact that modules may share genes under certain conditions.

Since VISDA applies hierarchical clustering, it indicates the similarities between the gene modules and provides flexibility in the resolutions/scales at which one identifies the modules. To get gene modules with bigger sizes, we can simply merge the clusters according to the hierarchy. In addition, the hierarchical relationship between the gene modules may provide additional biological insights, which is an advantage of VISDA over some parallel clustering methods like SOM and CLICK.

There is no theoretic barrier to use 3-D visualization and initialization, which will be a further improvement in VISDA. Two important things about using nonlinear mapping for visualization are: 1) a corresponding reverse mapping must be defined; 2) the distribution of projected data should not have large deviation from the Gaussian mixture assumption.

Subsequent work will investigate gene clustering at other nodes of the diagnostic tree, such as node 8. We will also compare the clustering results obtained at different nodes. We will measure the intersection of the clustering partitions, to identify conservative gene modules.

V. Conclusion

The incorporation of a hierarchical SFNM model, the complementary structure-preserving data projections, MDL model selection, and human interaction, enables VISDA to effectively detect gene clusters and to learn the hierarchical relationships between them. As a pilot study on the pathogenesis of muscular dystrophies at the gene module and gene regulation levels, we used IPA to analyze the gene modules discovered by VISDA under different phenotypic conditions and then developed some valuable insights to be further pursued in our future research.

Acknowledgment

This work is supported by the National Institutes of Health under grants CA109872, NS29525, CA096483, and caBIGTM. The muscular dystrophy dataset is provided by the Center for Genetic Medicine of the Children's National Medical Center.

References

[1] P. Tamayo, et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, pp. 2907-12, March 1999

[2] R. Sharan and R. Shamir, "CLICK: A clustering algorithm with applications to gene expression analysis," in *Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology*, 2000, pp. 307-16

[3] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Trans. on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24-45, 2004

[4] Y. Wang, L. Luo, M. T. Freedman, and S. Kung, "Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization," *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 625-36, 2000

[5] Z. Wang, et al., "Discriminatory mining of gene expression microarray data," *J. VLSI Signal Processing*, vol. 35, no. 3, pp. 255-72, 2003

[6] C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *IEEE Trans. Pattern Anal. Machine Intel*, vol. 20, pp. 282-93, 1998

[7] M. Bakay, et al., "Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration," *Brain*, vol. 192, no. 4, pp. 996-1013, 2006

[8] J. Rissanen, "Modeling by shortest data description," *Automatica*, Vol. 14, pp. 465-71, 1978

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. the Royal Statistical Society*, *Series B*, vol. 34, pp. 1-38, 1977

[10] http://www.ingenuity.com/