

Identifying Complex Biological Interactions based on Categorical Gene Expression Data

Ben Goertzel, Cassio Pennachin, Lúcio de Souza Coelho and Maurício Mudado

Abstract— A novel method, MUTIC (Model Utilization-based Clustering), is described for identifying complex interactions between genes or gene-categories based on gene expression data. The method deals with binary categorical data, which consists of a set of gene expression profiles divided into two biologically meaningful categories. It does not require data from multiple time points. Gene expression profiles are represented by feature vectors whose component features are either gene expression values, or averaged expression values corresponding to GO or PIR categories. A supervised learning algorithm (genetic programming) is used to learn an ensemble of classification models distinguishing the two categories based on the feature vectors corresponding to their members. Each feature is associated with a “model utilization vector,” which has an entry for each high-quality classification model found, indicating whether or not the feature was used in that model. These utilization vectors are then clustered using a variant of hierarchical clustering called Omniclust. The result is a set of model-utilization-based clusters, in which features are gathered together if they are often considered together by classification models – which may be because they’re co-expressed, or may be for subtler reasons involving multi-gene interactions. The MUTIC method is illustrated via applying it to a dataset regarding gene expression in human brains of various ages. Compared to traditional expression-based clustering, MUTIC yields clusters that have higher mathematical quality (in the sense of homogeneity and separation) and also yield novel insights into the underlying biological processes.

I. INTRODUCTION

A variety of methodologies for analyzing gene expression data have arisen in recent years, including but not limited to: identifying which genes are maximally differentiated between two categories; clustering genes based on coexpression across multiple samples or multiple experiments [1]-[8]; using supervised categorization algorithms to learn rules distinguishing two or more categories of gene expression profiles from each other [9]-[13]; and inference of genetic interaction networks from gene expression time series data [14]-[18]. These methodologies serve various purposes, such as induction of diagnostic models, qualitative understanding of the biological phenomena underlying a dataset, and identification of specific actors (e.g. genes, proteins) that may be involved in a certain biological phenomenon. In this paper we present a novel methodology for gene expression data analysis, whose goal is to identify those interactions

between genes, proteins, and biological processes that are most relevant to the phenotypic distinction underlying a given binary categorization of gene expression profiles.

Clustering is the most common tool for interaction identification. By determining which genes or gene-categories have expression-value profiles that cluster together across multiple samples or multiple experiments, one gets a picture of which genes are “associated” with each other. These associations do not usually have a clear interpretation, however, as co-expression can occur for a variety of reasons. Furthermore, many types of interactions are in principle not identifiable via directly clustering gene expression values. For instance, one won’t recognize ternary interactions wherein, say, C is only highly expressed when both A and B are highly expressed together.

The technique we describe here, MUTIC (Model Utilization-based Clustering), is oriented toward capturing interactions that ordinary expression-based clustering misses. The end result of MUTIC looks superficially similar to that of traditional gene expression clustering: one obtains a set of clusters (of genes or gene-categories), where the elements of a cluster are hypothesized to have a significant interrelationship. What is novel is that these clusters are not determined based on co-expression but via a more involved analysis. The semantics of the clusters is different: MUTIC clusters represent genes or gene-categories that are usefully considered in combination when formulating classification rules distinguishing one category of gene expression profiles from another. The elements of such a cluster may or may not be coexpressed across the set of gene expression profiles under analysis.

Here we describe the MUTIC method and then briefly discuss its application to a dataset regarding gene expression in human brain cells, collected in a study of the neurogenetics of aging [19]. In the context of this dataset, we review a number of potentially interesting biological interactions that the new method finds but traditional expression-based clustering misses. We also analyze homogeneity and separation properties of MUTIC clusters, coming to the conclusion that they possess significantly greater cluster quality than clusters found via traditional gene expression clustering.

II. THE MUTIC ALGORITHM

A. Data Requirements and Pre-processing

MUTIC deals with data which is categorical: i.e. one must start with a set of gene expression profiles belonging to one

Ben Goertzel (e-mail: bgoertzel@biomind.com), Cassio Pennachin (e-mail: cassio@biomind.com), Lúcio de Souza Coelho (e-mail: lucio@biomind.com) and Maurício Mudado (mauricio@biomind.com) work at Biomind LLC, Rockville, Maryland.

category, and a set belonging to another category. Also, the MUTIC algorithm assumes that the data supplied to it has already been pre-processed and normalized using standard algorithms (e.g. log transformation and z-score normalization).

As a preparation for MUTIC analysis, each gene expression profile in the dataset under analysis is associated with a numerical “feature vector.” The entries of this vector are either (normalized, transformed) gene expression values, or else values derived from these, each one corresponding to the average gene expression across a certain Gene Ontology or Protein Information Resource category [20]-[21]. We call these “enhanced feature vectors” [22]. These enhanced feature vectors are the inputs to the MUTIC process.

B. Supervised Classification of Gene Expression Data

Given a set of gene expression profiles divided into two categories, the MUTIC algorithm begins by learning an ensemble of classification models, each of which distinguishes the two categories using some learned rule. This classification-rule-learning step may be carried out using any of a variety of supervised categorization algorithms. Reference [11] provides a review and comparison of multiple learning techniques in several microarray datasets. Lately, support vector machines and ensemble methods, especially boosting, seem to be the most prevalent ones for microarray data categorization. In our work with MUTIC so far, we have used the genetic programming algorithm [23].

What is required in general is that the classification model learning algorithm has two properties:

- The ability to learn a variety of models of reasonably high classification accuracy.

- The ability to produce models in which it’s possible to state which of the features of the input vector (which genes; or in the case of enhanced feature vectors, which gene categories) are important for the model.

In the case of genetic programming, the classification portion of the algorithm is carried out by performing multiple independent learning runs, and picking the most accurate models from each run. The merit of GP from the perspective of MUTIC is that it tends to produce a number of qualitatively-different yet accurate models, which use different feature sets.

C. Ensemble-Based Classification Utility

Given an ensemble of classification models corresponding to a given dataset, we may then mine the model ensemble for information about the usefulness of each feature. In the MUTIC algorithm as currently implemented, we simply assume that all features in a given GP-learned model are equally important to that model. Based on experimentation on a number of datasets, we have found that sophisticated approaches don’t provide significant improvements over this simple technique, at least not for reasonably large model ensembles.

We then tabulate, for each feature, the percentage of the

classification models in which it is rated as important. This gives us a list of the features that are most useful for distinguishing the two categories – since they are frequently used as tools for building accurate classification models. The ordered list of most useful features may be subjected to qualitative biological relevance analysis.

We then construct a data structure called a Utility Profile. This is essentially an inversion of the list of classification utilities calculated for an ensemble of classification models. Suppose one has executed n categorization tasks over the same dataset D , thereby producing a set E of n classification models. E lets one calculate L , a set of n lists of important features, one for each model in E . Let F be the set of all important features present in any of the lists in L , i.e., F is the union of all lists in L . Now, for each element f in F a feature vector is constructed where feature corresponds to a model e in E . This is called the utilization vector associated with f . The entry of f ’s utilization vector corresponding to the model e contains the classification utility value of f in e .

Note that this technique implicitly assumes probabilistic independence between any two given models in the ensemble, an assumption that is unlikely to be accurate. As a consequence, our estimates of feature usefulness will tend to over-weight usefulness relative to models that cluster together in model-space. While there are solutions to this issue, they are very computationally expensive, and we have found that for reasonably large model ensembles the models seem to be relatively uniformly spread across the model ensemble, meaning that no strong biases tend to be introduced.

D. Omniclust

Once utilization vectors have been constructed for all the relevant features, the next and final step in the MUTIC methodology is to cluster the utilization vectors. One may use essentially any clustering algorithm here; after experimenting with a number of alternatives, we settled on a technique of our own construction called Omniclust, which is a simple variation on standard hierarchical clustering.

Generally we feel that hierarchical rather than partitioning based clustering is more appropriate in a MUTIC context, because one is principally looking for small sets of features that have strong interactions. Standard hierarchical clustering [2] does produce small clusters but at its lowest levels it can be prone to artifacts due to the arbitrary nature of the binary groupings it performs. For instance, if there is a natural grouping of three genes, standard binary hierarchical clustering won’t necessarily find it, but may instead either divide it among two or even three groupings of two; at best it will merge it into a grouping of four, together with another gene that isn’t as closely related to the other three. Omniclust follows the basic logic of hierarchical clustering but isn’t based on an arbitrary binarization. Other researchers have recently proposed alternative hierarchical clustering algorithms that also deviate from binary hierarchical modeling as well [24]-[25].

We now describe the Omniclust algorithm in a general

mathematical setting; the application to clustering utilization vectors will be apparent. Let $G=(V,E)$ be a non-directed, weighted graph where nodes in V are elements to be clustered and the edges in E are weighted by the similarity measurement between the nodes connected by them. That is, for any a,b in V and $e=\{a,b\}$ in E , $\text{weight}(e)=\text{similarity}(a,b)$. Then, the basic Omniclust step is:

Omiclust(G)

- 1) $S \leftarrow \{\}$ (**Initialize as empty the set of edges to be preserved.**)
- 2) **For each v in V do**
 - a) **Let $\text{edges}(v)$ be the set of all edges connecting v to other vertices.**
 - b) **Let s be the heaviest edge in $\text{edges}(v)$**
 - c) $S \leftarrow S + \{s\}$
- 3) $E \leftarrow S$ (**Deletes all edges that were not selected for preservation by any node inspection above. After this step, G will typically be partitioned in many subgraphs – called “clustlets” - in tree and line topologies.**)
- 4) **Let C be the set of connected subgraphs of G . (Defines the output set of all clustlets.)**
- 5) **Return C**

The clustlets themselves can then be used as nodes in a new graph that is then presented to Omniclust, and the process can be repeated again and again until Omniclust produces just one cluster, which will be the root of a hierarchical clustering based on graph-partitioning in each level. This process of successive clustering is described in the algorithm below:

HierarchicalOmiclust(G)

- 1) $V \leftarrow \text{Omiclust}(G)$
- 2) **while $|V| > 1$ do**
 - a) **Let E' be the set of weighted edges connecting all possible pairs of distinct clusters a,b in V such that each edge is weighted by an intercluster similarity measurement $s(a,b)$. It is also defined that $\text{weight}(a,b)=s(a,b)$.**
 - b) $G \leftarrow (V, E')$
 - c) $V \leftarrow \text{Omiclust}(G)$
- 3) **return V**

Although Omniclust is hierarchical, its first level, which contains the “clustlets,” offers a natural partition of the entities into clusters, as the size of clustlets is not fixed and is entirely defined by the similarity relations between the clustered entities. For sake of simplifying the biological interpretation of the results, all our analysis of clustering quality has been done considering clustlets only.

E. Choice of Clustering Similarity Metric

Omiclust, like most clustering algorithms, relies on an externally-defined similarity metric. For our work with MUTIC, we have chosen the cosine similarity measure. This choice was a consequence of the sparse nature of the

feature vectors in Utility Profiles produced by ensembles composed by Genetic Programming-evolved classification models. Any given model uses only a handful of features and therefore even a whole ensemble uses only a small subset of all available features in a dataset. In such a feature utilization scenario, the utility of a given gene or gene family for a given ensemble will be zero for most models, and therefore the corresponding vectors in the Utility Profile will be sparse. Cosine similarity is often used in other machine learning domains involving sparse vectors (such as text classification using word frequency vectors[26]) due to its capacity to compute meaningful similarity values in the face of severe sparseness.

III. RESULTS

A. Test Dataset

In this section we describe results obtained from applying the MUTIC methodology to the dataset described in [19], which reports microarray analysis of gene expression changes in post-mortem brain samples of frontal cortex from 30 individuals ranging in age from 26 to 106 years. Material acquisition and data preparation are described in [19] and will not be repeated here.

In [19] this dataset is analyzed in a conventional way. After looking for genes whose expression correlates significantly with age, clusters of genes that are up and down-regulated in aged and young individuals are identified. In a large subset of genes, negative correlation is found when comparing the gene expression from the group of “Young” individuals (less than 43 years old) versus “Old” ones (more than 74 years old). Many of these genes are related to synaptic function, neuronal plasticity, signal transduction, vesicular transport, protein metabolism, Ca^+ homeostasis, microtubule cytoskeleton, amino acid modification, hormones and immune response.

For our work here we have taken the subset of 21 individuals belonging to the categories “Young” or “Old,” and interpreted this as a binary categorization problem. This subset of the dataset given in [19] will be referred to as the “Aging Brain” dataset in the remainder of this paper. The goal of MUTIC, as applied to this dataset, is to determine sets of genes or gene categories whose interrelationships are important in the context of the aging of human brains.

B. Experimental and Analytical Setup

The Aging Brain dataset was used to produce Utility Profiles, via running a large number of differently-configured genetic programming based categorization processes to create a diverse classification model ensemble. The execution of the genetic programming algorithm was done using the Biomind ArrayGenius Software¹. In particular, we used the metatasking capability of ArrayGenius: the software, upon receiving the dataset as

¹ Available at ondemand.biomind.com

input, was instructed to run 1,000 GP processes with parameters selected randomly within specified ranges. Ranges used for parameter variation are detailed below (parameters not mentioned were left at their ArrayGenius default values):

--All combinations of use of direct and categorical features (see section above on feature vector enhancement) were allowed.

--In terms of GP-specific parameters, fitness function was varied across all available alternatives.

--The feature selection method used was Most Differentiated Features ranging from 10 to 1000 selected features in all tasks.

The Utility Profiles produced in this manner were then used as inputs for Omniclust clustering. This produces a set of feature clusters (where each feature can be a gene or a gene-category, as described above). These clusters may have a more general semantics than clusters formed from gene expression vectors directly using standard methods. In these utilization-based clusters, features are gathered together if classification models habitually found it useful to consider them together.

For sake of comparison, we also used Omniclust to perform clustering of the Aging Brain dataset in the traditional way, via simply clustering the feature vectors associated with the gene expression profiles.

In the remainder of this section, we compare the clustering results obtained by the two methods (MUTIC and traditional gene expression clustering) both quantitatively and qualitatively. The purpose of the quantitative comparison is to show that the clusters obtained using MUTIC are “better clusters” in a purely mathematical sense. The purpose of the qualitative comparison is to show that the clusters obtained using MUTIC yield novel biological insights not implicit in their traditional counterparts. A thorough qualitative exploration of the results on any one of the test datasets would be well beyond the scope of this paper. Here our focus is on the methodology and we will therefore restrict ourselves to a few observations regarding the unique qualitative insights provided by the MUTIC method.

C. Quantitative Comparison

Clustering is a qualitative data analysis method; there are no robust, commonly accepted, objective metrics for comparing different clustering algorithms to each other. [2] gives a comprehensive overview of contemporary clustering methods and a review of methods for comparing them to each other.

Choosing a variant of a standard technique, we have measured the quality of a clustering as the product homogeneity x separation. Homogeneity is calculated as $1/(1+A)$ where A is the average of the distances of all members of the cluster to their nearest cluster-mates. Separation is simply the minimum distance from any given member of the cluster to elements outside the cluster. These particular definitions of separation and homogeneity were

TABLE I
CLUSTERING QUALITIES

Clustering	Quality of 1 st cluster	Quality of 20 th cluster
Utilization-based clustering	0.4826	0.4516
Expression-based clustering	0.0045	0.0012
Expression-based clustering of top 3,913 most differentiated features	0.0065	0.0013
Expression-based clustering of top 3,913 most differentiated features using Average sparseness policy	0.138	0.054
Expression-based clustering of top 3,913 most differentiated features using Median sparseness policy	0.097	0.083
Expression-based clustering of top 3,913 most differentiated features using Custom sparseness policy	0.176	0.004

Clustering qualities for the approaches described in the text. Ordinal cluster numbers in the header refer to the clustering quality rank.

used in order to minimize the influence of the size of the cluster on its quality. (As we have observed empirically, using more traditional definitions of separation and homogeneity, e.g defining homogeneity as the average of all similarities between all members of a cluster, causes small clusters to habitually display a better quality than larger ones, which is an undesirable bias.)

If one straightforwardly compares MUTIC to plain expression-based clustering, according to this cluster quality metric, one finds that MUTIC produces dramatically clearer clusters, with roughly 100 times greater quality. This comparison however is somewhat unfair to the standard method, because the separation values are bound to be larger for MUTIC simply because it involves fewer features (only the ones that have nonzero model usage). Thus, to make a fairer comparison, we also tried standard expression-based clustering using a smaller set of features: only the N most-differentiated features, where differentiation was measured using the same categories used for supervised categorization, and N was chosen as the same number of features having nontrivial Utility Profiles. The results of this comparison are shown in Table 1. As we see, MUTIC still comes out far ahead here, with roughly 100 times higher cluster quality.

Another possible source of unfair comparisons could be the sparse nature of utility-based vectors as compared to gene expression vectors. In order to detect a potential unfair advantage based on sparseness, we applied three different sparseness policies to the gene expression vectors:

-- Average Policy: all values in a given feature vector below the average of those values were set to zero.

-- Median Policy: all values in a given feature vector below the average of those values were set to zero.

-- Custom Policy: in a generalization of the Median Policy, in this one all the lowest P % values in a given feature vector are set to zero. P was chosen as the average sparseness ratio (number of zero-ed dimensions over the total number of dimensions) in the utility-based data.

Using any one of these three sparseness policies raises the quality of the expression-based clustering to the same order

of magnitude as the utility-based clustering. Nevertheless, even the highest quality value (achieved using the Custom Policy) is roughly 1/3 of the quality obtained for utility-based clustering. Also, the quality differences between the 1st and 20th ranked clusters indicate a sharper decline of quality as rank increases when clustering sparsified expression vectors as opposed to utility vectors. It appears, therefore, that only part of the high quality of the utility based clusters is explained by the sparseness of the utility vectors. Even when this is accounted for, MUTIC provides substantially crisper clusters than expression based clustering.

We emphasize that our cluster quality assessment method was in no way engineered to favor the utilization-based clusters; and nor was the Omniclust method devised specifically to showcase utilization-based clustering, in fact it was devised for standard expression-based clustering and is used for this purpose within the Biominde ArrayGenius product. The essential result is that the clusters found via utilization-based clustering are drastically more clear and distinct than what traditional expression-based clustering yields.

D. Qualitative Comparison

Next, we present a brief qualitative analysis of the results obtained by applying MUTIC to the Aging Brain dataset. We review each of the top 5 MUTIC clusters in detail, commenting on the plausibility of each cluster as a nexus of gene and gene-category interactions, according to the knowledge contained in the current biological literature.

Cluster #1 is composed of the features:

- ankyrin repeat domain 12 (ANK)
- MAX interactor 1 (MXI1)
- Mesenchyme homeo box 2 (MEOX2)
- MADS box transcription enhancer factor 2 (MEF2)
- Homo sapiens mitogen-activated protein kinase 14 (MAPK14 or MXI2 or p38)

This cluster interrelates several proteins found in the brain, with well-known brain-specific roles. MEF2, MXI1, MAPK14 and MEOX2 are all related to cell development and differentiation [27]-[31].

Interestingly all these features are either directly DNA-binding proteins (MEF2, MEOX2) or they indirectly interact with DNA-binding proteins: MXI1 and MAPK14/p38/MXI2 with MAX protein [32]; MAPK14/p38/MXI2 with MEF2 [33]; ANK with P160 cofactors/activators [34]. Therefore, they are directly related to regulation of gene expression. Furthermore, both MEF2 and MAPK14 have roles in apoptosis and cell development [35]-[36]. Moreover, since MAPK14 has a strong role in Alzheimer's Disease (AD) [37], we might investigate whether any of the other genes can also be related to this condition.

Next, Cluster #2 is composed of the features:

- KIAA0493 protein.
- Homo sapiens gamma-butyrobetaine hydroxylase 1 (BBOX1).
- 34273_at - Homo sapiens regulator of G-protein

signalling 4 (RGS4).

- Sterol carrier protein 2. (SCP-2)
- Leukocyte immunoglobulin-like receptor. (MHC)
- Homo sapiens 1,3-galactosyltransferase (B3GALT3).
- NM_005613 - Regulator of G-protein signalling 4. (RGS4)
- Hypothetical protein MGC35048.
- Homo sapiens histidine triad nucleotide binding protein 1 (HINT1).
- Homo sapiens protein tyrosine phosphatase (PTPRO).
- Homo sapiens S-phase kinase-associated (SKP1A).
- Homo sapiens diacylglycerol kinase (DGKB).
- Homo sapiens syntrophin alpha 1 (SNTA1).
- Homo sapiens cyclin-dependent kinase 5 (CDK5).
- Phosphoserine phosphatase (PPHOS).
- KIAA1240 protein.

This is a large cluster, but apparently its genes are associated with 3 major functions/characteristics: neuronal development and plasticity: MHC, UDP-Gal, PTPRO, SKP1A, DGKB, SNTA-1, CDK5, PPHOS [38]-[40] SKP1A, DGKB, SNTA-1, CDK5, PPHOS – [39],[41]-[45], brain/psychiatric disorders and diseases like AD, schizophrenia, bipolar disorder and brain cancer: RGS4, UDP-Gal, HINT1, DGKB, CDK5, PPHOS [46]-[49] DGKB, CDK5, PPHOS – [49]-[51], and signal transduction pathways molecules like ATPases, kinases, phosphatases and G-protein regulation: RGS4, SCP-2, HINT1, PTPRO, DGKB, CDK5, PPHOS.

Another interesting feature present here is lipid and ganglioside transport through membranes (BBOX1, SCP2, UDP-Gal). Gangliosides (UDP-Gal) are related to neuronal plasticity [52] and are possibly a substrate for SCP-2, as one of their functions is to mediate the transport of gangliosides through membranes.

Two identical features with same function, but different IDs (34273_at and NM_005613, both related to RGS4), were gathered together here (a case where MUTIC joins together two features that would also be joined by traditional clustering). RGS4 also inhibits the p38 (MAPK14 - a feature from cluster 1) signal transduction pathway [53]. And, as this cluster contains predicted and hypothetical proteins (KIAA0493, KIAA1240 and Hypothetical protein MGC35048), it provides suggestions for potential wet lab experiments concerning the role of these proteins.

Cluster #3 is composed of the features:

- Flap structure-specific endonuclease 1(RAD2).
- CAP, adenylate cyclase-associated protein.
- Cell division cycle 2-like (CDC2L5).
- Jun D proto-oncogene (JUND).
- COP9 constitutive photomorphogenic homolog subunit 2.
- Cut-like 2 (CUX2)
- Homo sapiens actin alpha 1 (ACTA1).

Five out of 7 features in this cluster have known, indirect interrelationships (CDCL5, JUND, COP9, CUX2 and ACTA1). Of these, 3 have been related to the cell cycle:

CDLC5, JUND and ACTA1 [54]-[56]. COP9 is also indirectly related to the cell cycle as it is known to be a part of a complex that regulates JUND [57]. Furthermore, COP9, CDLC5 and JUND have similar molecular functions as they all have protein kinase functions. CUX2 is a DNA-binding molecule (transcription factor) like JUND and is related to brain differentiation [58]-[59]. We may ask if these cell cycle features are somehow connected to the Huntington Disease by the RAD2 feature [60] and to cerebrovascular injury by the CAP feature [61].

Cluster #4 is composed by the features:

- Mitogen-activated protein kinase kinase 1 (MAP2K1)
- MRPL28 - mitochondrial ribosomal protein L28
- Homo sapiens reticulon 4 (RTN4)
- Translocase of inner mitochondrial membrane 17 (TIMM17A)

It has been postulated that the mitochondria have several roles in ageing, specially in the brain and AD [62]. In this cluster we find two features related to mitochondria: MRPL28 and TIMM17A. The last is involved in neurodegeneration [63] and the former has been related to some kinds of cancer [64]. MAP2K1 is also related to AD and has roles in neuronal plasticity and brain development [65-67]. RTN4 is related to neuronal plasticity [68]-[69].

Finally, Cluster #5 is composed of the features:

- DnaJ (Hsp40) homolog (DNAJB4)
- Homo sapiens metallothionein 1B (MT1B).
- S-adenosylhomocysteine hydrolase-like 1 (AHCYL1)
- Homo sapiens EGF-containing fibulin-like extracellular matrix protein 2 (EFEMP2).
- Iduronate 2-sulfatase (IDS)
- IQ motif and Sec7 domain 1 (IQSEC1).
- Homo sapiens adducin 3 (ADD3).
- Homo sapiens aldolase A (ALDOA).

This cluster has two closely related features with function of neuroprotection in neurodegenerative diseases: DNAJB4 and MT1B [70-73] though no experimental report had linked them before. Some features are related to diseases like the Huntington Syndrome and epilepsy: IDS and AHCYL1 [74]-[75]. Perhaps a possible connection between the two neuroprotective genes related to epilepsy and Huntington Syndrome symptoms is worth exploring.

IV. DISCUSSION

We have presented a novel analytical method, Model Utilization-based Clustering or MUTIC, and explored its behavior via discussing the results of its application to a gene expression dataset pertinent to the neurogenetics of human aging. The method has shown itself able to produce clusters with high mathematical significance, and also to identify interesting inter-gene and inter-process interactions that were not identified via standard expression-based clustering on the same dataset.

Like standard expression-based clustering, MUTIC is ultimately a method of qualitative data analysis, and therefore the evaluation of the method is not a simple thing. The true test of the method will be whether, when applied

across a wide variety of datasets and interpreted by researchers familiar with those datasets and their biological contexts, the method is successful at directing researchers toward useful and novel interpretations of their data. However, there is also an objective component to the advantage of the present approach over traditional clustering, in that there are some types of interrelationship that utilization-based clustering can capture which traditional expression-based clustering is mathematically unable to capture. And the results we have presented on comparative clustering quality show that the MUTIC clusters have more intrinsic mathematical validity than clusters found via the traditional approach, at least on the dataset considered here.

While we have dealt only with static gene expression data in this work, the method can be applied to time series data, when available, and we intend to do so in the future, along with carrying out more extensive applications to other gene expression datasets.

Finally, it should be noted that the MUTIC algorithm itself is not restricted to gene expression data, but may be of much more general value in a variety of different domains. It is potentially applicable to any dataset that is meaningfully treatable as categorical and that displays complex inter-feature interactions.

REFERENCES

- [1] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *J Comput Biol* 6: 281-297, 1999.
- [2] J. Dopazo and F. Azuaje, *Data analysis and visualization in genomics and proteomics*. John Wiley, Chichester, West Sussex; Hoboken, NJ, 2005.
- [3] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A* 95: 14863-14868, 1998.
- [4] P.E. Neiman, A. Ruddell, C. Jasoni, G. Loring, S.J. Thomas, K.A. Brandvold, R. Lee, J. Burnside and J. Delrow. "Analysis of gene expression during myc oncogene-induced lymphomagenesis in the bursa of Fabricius," *Proc Natl Acad Sci U S A* 98: 6378-6383, 2001.
- [5] R. Sharan and R. Shamir. "CLICK: a clustering algorithm with applications to gene expression analysis," *Proc Int Conf Intell Syst Mol Biol* 8: 307-316, 2000.
- [6] R. Sharan, R. Elkon, and R. Shamir, "Cluster analysis and its applications to gene expression data", in *Ernst Schering workshop on Bioinformatics and Genome Analysis*. Springer Verlag, 2001.
- [7] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol Biol Cell* 9: 3273-3297, 1998.
- [8] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub. "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc Natl Acad Sci U S A* 96: 2907-2912, 1999.
- [9] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler. "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc Natl Acad Sci U S A* 97: 262-267, 2000.
- [10] J.H. Cho, D. Lee, J.H. Park, and I.B. Lee, "Gene selection and classification from microarray data using kernel machine," *FEBS Lett* 571: 93-98, 2004.
- [11] S. Dudoit, J. Fridlyand and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J Am Stat Assoc* 97: 77-87, 2002.

- [12] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science* 286: 531-537, 1999.
- [13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning* 46: 389-422, 2002.
- [14] F. Markovetz. *A bibliography on learning causal networks of gene interactions*. 2004
- [15] F. Markovetz and R. Spang, "Reconstructing gene regulation networks from passive observations and active interventions," in *7th Ann Intl Conf Res Comput Molec Biol (RECOMB)*, 2003.
- [16] I. Nachman, A. Regev, and N. Friedman. "Inferring quantitative models of regulatory networks from expression data," *Bioinformatics* 20 Suppl 1: I248-I256, 2004
- [17] F. Sohler, D. Hanisch and R. Zimmer. "New methods for joint analysis of biological networks and expression data," *Bioinformatics* 20: 1517-1521, 2004.
- [18] J.P. Vert and M. Kanehisa. "Extracting active pathways from gene expression data," *Bioinformatics* 19 Suppl 2: II238-II244, 2003.
- [19] T. Lu, Y. Pan, S.Y. Kao, C. Li, I. Kohane, J. Chan, and B.A. Yankner. "Gene regulation and DNA damage in the ageing human brain," *Nature* 429: 883-891, 2004.
- [20] 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, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet* 25: 25-29, 2000.
- [21] C.H. Wu, L.S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R.S. Ledley, B.E. Suzek, C.R. Vinayaka, J. Zhang, and W.C. Barker. "The Protein Information Resource," *Nucleic Acids Res* 31: 345-347, 2003.
- [22] C.. Pennachin, L.S.. Coelho, I. Goertzel, M. Queiroz, F. Prosdociami, F. Lobo, and B. Goertzel. "Knowledge-Guided Analysis of Gene Expression Data Using Genetic Programming, Support Vector Machines, and the Gene Ontology and PIR Databases," submitted to *JBCB*, 2005.
- [23] J.R. Koza. *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge, Mass. 1992.
- [24] Z. Bar-Joseph, E.D. Demaine, D.K. Gifford, N. Srebro, A.M. Hamel, and T.S. Jaakkola, "K-ary clustering with optimal leaf ordering for gene expression data," *Bioinformatics* 19: 1070-1078, 2003.
- [25] E. Segal and D. Koller. Probabilistic hierarchical clustering for biological data. In *6th Ann Intl Conf Res Comput Molec Biol (RECOMB)*, 2002.
- [26] C.J. Van Rijsbergen. *Information retrieval*. Butterworths, London; Boston, 1979.
- [27] R.L. Eckert, T. Efimova, S. Balasubramanian, J.F. Crish, F. Bone, and S. Dashti, "p38 Mitogen-activated protein kinases on the body surface--a function for p38 delta," *J Invest Dermatol* 120: 823-828, 2003.
- [28] D.H. Gorski and A.J. Leal. "Inhibition of endothelial cell activation by the homeobox gene Gax," *J Surg Res* 111: 91-99, 2003.
- [29] X. Lin, S. Shah and R.F. Balleit. "The expression of MEF2 genes is implicated in CNS neuronal differentiation," *Brain Res Mol Brain Res* 42: 307-316, 1996.
- [30] I. Manni, P. Tunici, N. Cirenei, R. Albarosa, B.M. Colombo, L. Roz, A. Sacchi, G. Piaggio, and G. Finocchiaro, "Mxi1 inhibits the proliferation of U87 glioma cells through down-regulation of cyclin B1 gene expression," *Br J Cancer* 86: 477-484, 2002.
- [31] N. Schreiber-Agus, Y. Meng, T. Hoang, H. Hou, Jr., K. Chen, R. Greenberg, C. Cordon-Cardo, H.W. Lee, and R.A. DePinho. "Role of Mxi1 in ageing organ systems and the regulation of normal and neoplastic growth," *Nature* 393: 483-487, 1998.
- [32] A.S. Zervos, L. Faccio, J.P. Gatto, J.M. Kyriakis, and R. Brent, "Mxi2, a mitogen-activated protein kinase that recognizes and phosphorylates Max protein," *Proc Natl Acad Sci U S A* 92: 10531-10534, 1995.
- [33] M. Zhao, L. New, V.V. Kravchenko, Y. Kato, H. Gram, F. di Padova, E.N. Olson, R.J. Ulevitch, and J. Han, "Regulation of the MEF2 family of transcription factors by p38," *Mol Cell Biol* 19: 21-30, 1999.
- [34] O.C. Meijer, P.J. Steenbergen, and E.R. De Kloet, "Differential expression and regional distribution of steroid receptor coactivators SRC-1 and SRC-2 in brain and pituitary," *Endocrinology* 141: 2192-2199, 2000.
- [35] J.L. Kummer, P.K. Rao, and K.A. Heidenreich, "Apoptosis induced by withdrawal of trophic factors is mediated by p38 mitogen-activated protein kinase," *J Biol Chem* 272: 20490-20494, 1997.
- [36] D.A. Linseman, B.J. Cornejo, S.S. Le, M.K. Meintzer, T.A. Laessig, R.J. Bouchard, and K.A. Heidenreich, "A myocyte enhancer factor 2D (MEF2D) kinase activated during neuronal apoptosis is a novel target inhibited by lithium," *J Neurochem* 85: 1488-1499, 2003.
- [37] C.S. Piao, Y. Che, P.L. Han, and J.K. Lee, "Delayed and differential induction of p38 MAPK isoforms in microglia and astrocytes in the brain after transient global ischemia," *Brain Res Mol Brain Res* 107: 137-144, 2002.
- [38] N. Adachi, M. Oyasu, T. Taniguchi, Y. Yamaguchi, R. Takenaka, Y. Shirai, and N. Saito, "Immunocytochemical localization of a neuron-specific diacylglycerol kinase beta and gamma in the developing rat brain," *Mol Brain Res*, 2005.
- [39] Y.T. Bryceson, J.A. Foster, S.P. Kuppusamy, M. Herkenham, and E.O. Long, "Expression of a killer cell receptor-like gene in plastic regions of the central nervous system," *J Neuroimmunol* 161: 177-182, 2005.
- [40] P.J. Beltran, J.L. Bixby, and B.A. Masters, "Expression of PTPRO during mouse development suggests involvement in axonogenesis and differentiation of NT-3 and NGF-dependent neurons," *J Comp Neurol* 456: 384-395, 2003.
- [41] J.C. Cruzf and L.H. Tsai, "A Jekyll and Hyde kinase: roles for Cdk5 in brain development and disease," *Curr Opin Neurobiol* 14: 390-394, 2004.
- [42] D.C. Gorecki, H. Abdulrazzak, K. Lukasiuk, and E.A. Barnard, "Differential expression of syntrophins and analysis of alternatively spliced dystrophin transcripts in the mouse brain," *Eur J Neurosci* 9: 965-976, 1997.
- [43] L. Stepanek, Q.L. Sun, J. Wang, C. Wang, and J.L. Bixby, "CRYP-2/cPTPRO is a "neurite inhibitory repulsive guidance cue for retinal neurons in vitro," *J Cell Biol* 154: 867-878, 2001.
- [44] E. Uro-Coste, C. Fonta, F. Hatey, E. Perret, M.B. Delisle, D. Caput, and M. Imbert, Expression of SKP1 mRNA and protein in rat brain during postnatal development," *Neuroreport* 8: 1675-1678, 1997.
- [45] S. Verleysdonk and B. Hamprecht, "Synthesis and release of L-serine by rat astroglia-rich primary cultures," *Glia* 30: 19-26, 2000.
- [46] K. Goto and H. Kondo, "Diacylglycerol kinase in the central nervous system--molecular heterogeneity and gene expression," *Chem Phys Lipids* 98: 109-117, 1999.
- [47] K. Goto and H. Kondo, "Functional implications of the diacylglycerol kinase family," *Adv Enzyme Regul* 44: 187-199, 2004.
- [48] E.A. Monaco, 3rd. 2004, Recent evidence regarding a role for Cdk5 dysregulation in Alzheimer's disease," *Curr Alzheimer Res* 1: 33-38, 2004.
- [49] K.M. Prasad, K.V. Chowdari, V.L. Nimgaonkar, M.E. Talkowski, D.A. Lewis, and M.S. Keshavan, "Genetic polymorphisms of the RGS4 and dorsolateral prefrontal cortex morphology among first episode schizophrenia patients," *Mol Psychiatry* 10: 213-219, 2005.
- [50] N. Tsuda, Y. Nonaka, S. Shichijo, A. Yamada, M. Ito, Y. Maeda, M. Harada, T. Kamura, and K. Itoh, "UDP-Gal: betaGlcNAc beta1, 3-galactosyltransferase, polypeptide 3 (GALT3) is a tumour antigen recognised by HLA-A2-restricted cytotoxic T lymphocytes from patients with brain tumour," *Br J Cancer* 87: 1006-1012, 2002.
- [51] M.P. Vawter, J.M. Crook, T.M. Hyde, J.E. Kleinman, D.R. Weinberger, K.G. Becker, and W.J. Freed, "Microarray analysis of gene expression in the prefrontal cortex in schizophrenia: a preliminary study," *Schizophr Res* 58: 11-20, 2002.
- [52] R. Mendez-Otero and L.A. Cavalcante, "Functional role of gangliosides in neuronal motility," *Prog Mol Subcell Biol* 32: 97-124, 2003.
- [53] A.R. Albig and W.P. Schiemann, "Identification and characterization of regulator of G protein signaling 4 (RGS4) as a novel inhibitor of tubulogenesis: RGS4 inhibits mitogen-activated protein kinases and vascular endothelial growth factor signaling," *Mol Biol Cell* 16: 609-625, 2005.
- [54] A. Forer, "Does actin produce the force that moves a chromosome to the pole during anaphase?" *Can J Biochem Cell Biol* 63: 585-598, 1985.

- [55] D. Lallemand, G. Spyrou, M. Yaniv, and C.M. Pfarr, "Variations in Jun and Fos protein expression and AP-1 activity in cycling, resting and stimulated fibroblasts," *Oncogene* 14: 819-830, 1997.
- [56] F. Marques, J.L. Moreau, G. Peaucellier, J.C. Lozano, P. Schatt, A. Picard, I. Callebaut, E. Perret, and A.M. Genevriere, "A new subfamily of high molecular mass CDC2-related kinases with PITAI/VRE motifs," *Biochem Biophys Res Commun* 279: 832-837, 2000.
- [57] D.A. Chamovitz and D. Segal. "JAB1/CSN5 and the COP9 signalosome. A complex situation," *EMBO Rep* 2: 96-101, 2001.
- [58] M. Nieto, E.S. Monuki, H. Tang, J. Imitola, N. Haubst, S.J. Khoury, J. Cunningham, M. Gotz, and C.A. Walsh, "Expression of Cux-1 and Cux-2 in the subventricular zone and upper layers II-IV of the cerebral cortex," *J Comp Neurol* 479: 168-180, 2004.
- [59] C. Zimmer, M.C. Tiveron, R. Bodmer, and H. Cremer, "Dynamics of Cux2 expression suggests that an early pool of SVZ precursors is fated to become upper cortical layer neurons," *Cereb Cortex* 14: 1408-1420, 2004.
- [60] B. Shen, P. Singh, R. Liu, J. Qiu, L. Zheng, L.D. Finger, and S. Alas, "Multiple but dissectible functions of FEN-1 nucleases in nucleic acid processing, genome stability and diseases," *Bioessays* 27: 717-729, 2005.
- [61] M. Fornage, M.W. Swank, E. Boerwinkle, and P.A. Doris, "Gene expression profiling and functional proteomic analysis reveal perturbed kinase-mediated signaling in genetic stroke susceptibility," *Physiol Genomics* 15: 75-83, 2003.
- [62] S. Melov, "Modeling mitochondrial function in aging neurons," *Trends Neurosci* 27: 601-606, 2004.
- [63] M.F. Bauer and W. Neupert, "Import of proteins into mitochondria: a novel pathomechanism for progressive neurodegeneration," *J Inherit Metab Dis* 24: 166-180, 2001.
- [64] P.F. Robbins, M. el-Gamil, Y.F. Li, S.L. Topalian, L. Rivoltini, K. Sakaguchi, E. Appella, Y. Kawakami, and S.A. Rosenberg, "Cloning of a new gene encoding an antigen recognized by melanoma-specific HLA-A24-restricted tumor-infiltrating lymphocytes," *J Immunol* 154: 5944-5950, 1995.
- [65] R.J. Kelleher 3rd, A. Govindarajan, H.Y. Jung, H. Kang, and S. Tonegawa, "Translational control by MAPK signaling in long-term synaptic plasticity and memory," *Cell* 116: 467-479, 2004.
- [66] J.J. Pei, H. Braak, W.L. An, B. Winblad, R.F. Cowburn, K. Iqbal, and I. Grundke-Iqbal, "Up-regulation of mitogen-activated protein kinases ERK1/2 and MEK1/2 is associated with the progression of neurofibrillary degeneration in Alzheimer's disease," *Brain Res Mol Brain Res* 109: 45-55, 2002.
- [67] D. Refojo, C. Echenique, M.B. Muller, J.M. Reul, J.M. Deussing, W. Wurst, I. Sillaber, M. Paez-Pereda, F. Holsboer, and E. Arzt, "Corticotropin-releasing hormone activates ERK1/2 MAPK in specific brain areas," *Proc Natl Acad Sci U S A* 102: 6183-6188, 2005.
- [68] C.E. Ng and B.L. Tang, "Nogos and the Nogo-66 receptor: factors inhibiting CNS neuron regeneration," *J Neurosci Res* 67: 559-565, 2002.
- [69] R. Prinjha, S.E. Moore, M. Vinson, S. Blake, R. Morrow, G. Christie, D. Michalovich, D.L. Simmons, and F.S. Walsh, "Inhibitor of neurite outgrowth in humans," *Nature* 403: 383-384, 2000.
- [70] Dittmann, J., S.J. Fung, J.C. Vickers, M.I. Chuah, R.S. Chung, and A.K. West, "Metallothionein biology in the ageing and neurodegenerative brain," *Neurotox Res* 7: 87-93, 2005.
- [71] Ebadi, M., H. Brown-Borg, H. El Refaey, B.B. Singh, S. Garrett, S. Shavali, and S.K. Sharma, "Metallothionein-mediated neuroprotection in genetically engineered mouse models of Parkinson's disease," *Brain Res Mol Brain Res* 134: 67-75, 2005.
- [72] Ohtsuka, K. and T. Suzuki, "Roles of molecular chaperones in the nervous system", *Brain Res Bull* 53: 141-146, 2000.
- [73] Tamura, S., H. Kinouchi, K. Izaki, A. Okubo, T. Sugawara, H. Kunizuka, and K. Mizoi, "Induction of heat shock protein 40 and GrpE mRNAs following transient focal cerebral ischemia in the rat", *Brain Res* 960: 277-281, 2003.
- [74] Daniele, A., R. Tomanin, G.R. Villani, F. Zacchello, M. Scarpa, and P. Di Natale, "Uptake of recombinant iduronate-2-sulfatase into neuronal and glial cells in vitro", *Biochim Biophys Acta* 1588: 203-209, 2002.
- [75] Mita, T., I. Kawazu, H. Hirano, O. Ohmori, N. Janjua, and K. Shibata "El mice epilepsy shows genetic polymorphism for S-Adenosyl-L-homocysteine hydrolase", *Neurochem Int* 38: 349-357, 2001.