Clustering Gene Expression Data via Mining Ensembles Of Classification Rules Evolved Using MOSES

Moshe Looks

Department of Computer Science and Engineering
Washington University
Saint Louis, MO 63130, USA
moshe@metacog.org

Ben Goertzel, Lucio de Souza Coelho,
Mauricio Mudado, Cassio Pennachin
Biomind LLC
1405 Bernerd Place
Rockville, MD 20851, USA
ben@goertzel.org,
{lucio, mauricio, cassio}@vettalabs.com

ABSTRACT

A novel approach, model-based clustering, is described for identifying complex interactions between genes or gene-categories based on static gene expression data. The approach deals with categorical data, which consists of a set of gene expression profiles belonging to one category, and a set belonging to another category. An evolutionary algorithm (Meta-Optimizing Semantic Evolutionary Search, or MOSES) is used to learn an ensemble of classification models distinguishing the two categories, based on inputs that are features corresponding to gene expression values. Each feature is associated with a model-based vector, which encodes quantitative information regarding the utilization of the feature across the ensembles of models. Two different ways of constructing these vectors are explored. These model-based vectors are then clustered using a variant of hierarchical clustering called Omniclust. The result is a set of model-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 method is illustrated by applying it to two datasets regarding human gene expression, one drawn from brain cells and pertinent to the neurogenetics of aging, and the other drawn from blood cells and relating to differentiating between types of lymphoma. We find that, compared to traditional expression-based clustering, the new method often yields clusters that have higher mathematical quality (in the sense of homogeneity and separation) and also yield novel and meaningful insights into the underlying biological processes.

Categories and Subject Descriptors

I.2.2 [**Artificial Intelligence**]: Automatic Programming – Program synthesis

General Terms

Algorithms, Design, Experimentation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GECCO'07, July 7-11, 2007, London, England, United Kingdom.
Copyright 2007 ACM 978-1-59593-697-4/07/0007...$5.00.

Keywords

Empirical Study, Heuristics, Optimization, Representations

1. INTRODUCTION

A variety of methodologies for analyzing gene expression data have emerged 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 [3, 8, 10, 21, 23, 24, 25, 30], the application of supervised categorization algorithms to learn rules distinguishing two or more categories of gene expression profiles from each other [6, 7, 9, 12, 13], and the inference of genetic interaction networks from gene expression time series data [17, 18, 20, 27, 32]. These methodologies serve different purposes, such as induction of diagnostic models, qualitative understanding of the biological phenomena underlying a dataset, and the search for specific actors (e.g. genes, proteins) that may be involved in a biological phenomenon. This paper discusses a methodology that aims to identify relevant interactions between genes, proteins, and biological processes, allowing for a qualitative understanding of these interactions in the context of a microarray dataset.

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 have a clear biological interpretation, as co-expression can occur for a variety of reasons. Furthermore, many types of interactions will not be identified by this approach. For instance, one won't recognize ternary interactions wherein, say, C is only highly expressed when A and B are highly expressed together.

We present here a novel technique, model-based clustering, for identifying interactions from microarray gene expression data, which captures interactions that ordinary expression-based clustering misses. The end result of the analysis is familiar to bioinformaticians: a clustering of features (representing genes or gene-categories) that seem to have a significant interrelationship. What is novel is that these clusters are not determined based on co-expression but via more involved analysis.

In this paper we describe two variants of the model-based clustering method: Model-Based Utilization Clustering (MUTIC) and Model Based Role Analysis (MOBRA). We also discuss its application to two datasets: one consisting of gene expression

levels in human brain cells, collected in a study of the neurogenetics of aging; and one containing gene expression of human blood cells affected by different kinds of lymphoma. We discuss the relevant biological interactions that the new method finds for these datasets. We also consider the homogeneity and separation properties of clusters found via model-based clustering, concluding that they are of high quality in the examples studied.

2. METHODOLOGY

2.1 Outline

Model-based clustering has the following steps and requirements:

- *Data input:* the original data submitted to model-based clustering has to be categorical, since classification models are the entities actually used to produce the intermediate, model-based vectors used in clustering.
- *Classification:* data defined by the previous step is then submitted to a supervised learning algorithm, or classification method. A classification method employed by model-based clustering has to be able to produce a large number of different classification models. More specifically, the models should either use different feature subsets for performing classification, or be able to provide some metric for measuring the utility of a feature in classification.
- *Model-based Transformation:* for each feature in the dataset, a *model-based vector* is created, based on the utilization (or utility) of the feature across all models produced in the previous step (two possible ways of computing these vectors are described in Section 2.4).
- *Clustering:* a clustering algorithm is then applied to the transformed data (model-based vectors) generated in the previous step. It is expected that clusters produced that way will be able to give interesting insights into the inter-relationships among features concerning the differences between the dataset categories.

In the following sections, we will describe the datasets, classification method, model-based transformations and clustering algorithm used in the examples of model-based clustering to be presented.

2.2 Datasets

Two datasets were used in order to explore and validate the model-based clustering methodology:

Aging brain: this dataset is taken from [16], a 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. After looking for genes which expression correlates significantly with age, clusters of genes that are up and down-regulated in aged and young individuals were found. A negative correlation when comparing the gene expression from the group of young individuals (less than 42 years old) versus aged ones (more than 72 years old) was found in a large subset of genes, mostly related with synaptic function, neuronal plasticity, signal transduction, vesicular transport, protein metabolism, Ca+ homeostasis, microtubule cytosqueleton, aminoacid modification, hormones and immune response. This subset of 19 individuals belonging to the categories "Young" and "Old" comprises the so-called Aging Brain dataset used in the present study.

Lymphoma: this dataset is taken from [27]. It contains 58 cases of diffuse large B-cell lymphoma, and 19 cases of follicular lymphoma (77 total). It is interesting to stress that this dataset poses a somewhat different problem than the previous one, from a biological standpoint. While a dataset analyzing old versus new subjects is expected to find marked, sharper differences, the distinction between two types of the same condition (lymphoma) is expected to be subtler.

2.3 MOSES

Meta-optimizing semantic evolutionary search (MOSES) [15] is a recently developed program evolution system distinguished by two key mechanisms: (1) exploiting semantics (what programs actually mean) to restrict and direct search; and (2) limiting the recombination of programs to occur within bounded subspaces (constructed on the basis of program semantics). This has been shown to lead to superior performance and scalability in comparison to current purely syntactic techniques (local search, genetic programming, etc.). Furthermore, the evolved programs do not suffer from any kind of "bloating", and are generally quite comprehensible. This is of particular interest for applications such as microarray analysis, where it is useful to know not only that a method achieves good performance, but to understand how.

A detailed description of MOSES is beyond the scope of this paper. In the context of model-based clustering, however, it suffices to mention that MOSES produces an extensive collection of different classification models. Structurally and functionally, these are the same as GP models (program trees), though with a strong tendency for parsimony (small sizes). Feature utilization is obviously different among the models in a single ensemble; that is, the feature set used by each model will tend to be different from the others, though with some overlap.

The instance of MOSES used here works with logical operators only; the datasets were therefore discretized into Boolean values. This discretization was performed by taking the median value of a given feature in the dataset was a threshold, and assigning *false* to values below the threshold, *true* otherwise. On both datasets, MOSES was run with a total of 50 Boolean features corresponding to the most-differentiated genes. Ten independent runs with 10-fold cross-validation were used, with a total of 100,000 evaluations per run. In the end the best models produced – those with highest fitness and smallest size – were used for model-based clustering.

2.4 Model-Based Transformation

Two strategies of mapping vectors from classification models to features were devised here.

The first one is called Model-Based Utility Clustering, or MUTIC for short. A version of MUTIC, using ensemble of GP models, is described in [11]. In the specific setting of MOSES, a MUTIC vector for a given feature f in the dataset is generated as follows. Let $M=\{m_1, m_2, m_3,...,m_n\}$ be the models generated by the application of MOSES to the data. A MUTIC vector is defined by

$$V_{MUTIC}(f) = [u(f,m_1), u(f,m_2), u(f,m_3), ..., u(f,m_n)],$$

where $u(f,m)$ returns 1 if feature f is used by model m, 0 otherwise.

The other model-based vector mapping approach used is called Model-Based Role Analysis, or MOBRA. In order to describe how MOBRA operates, we consider again the set of features $F=\{f_1, f_2, f_3, ..., f_d\}$ in the dataset. A MOBRA model-based vector for a given feature f_i is then defined as

$$V_{MOBRA}(f_i) = \{c(f_i,f_1), c(f_i,f_2), c(f_i,f_3), ..., c(f_i,f_d)\},$$

where $c(f_i, f_j)$ returns 1 if f_i and f_j co-occur in at least one model in M, and 0 otherwise.

These different mappings can be thought of as "answering" two different questions about feature usage in classification models. MOBRA tries to answer the question of which features tend to play similar roles to each other, in the dataset in question. In the context of the datasets used here, of course, this refers to a biological role. MUTIC tries to answer the question of which features may be inter-related in some non-trivial way to produce the categorical outcomes observed in the dataset. This is particularly interesting for gene expression analysis, considering that genes form self-regulatory networks with complex, non-obvious gene-gene interactions. And of course, MOBRA and MUTIC are just two possible model-based mappings; other questions may lead to new model-based mappings.

2.5 Clustering

Once the model-based vectors have been constructed for all the relevant features, the final step in the model-based clustering 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 [11], which is a simple variation on the standard hierarchical clustering algorithm.

While choices between clustering algorithms are largely qualitative, our choice of Omniclust perhaps merits brief discussion. Generally we feel that hierarchical rather than partitioning based clustering is more appropriate in a utilization-based-clustering context, because one is principally looking for small sets of features that have strong interactions. Standard hierarchical clustering as used in bioinformatics [8] 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 recent hierarchical clustering algorithms seem to deviate from binary hierarchical modeling as well [2], [22]. The less arbitrary hierarchy qualitatively seems to display fewer odd artifacts for very small clusters, which are the ones of most interest from a utilization-based clustering perspective.

We now describe the first level of the Omniclust algorithm, in a general mathematical setting from which 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, $weight(e)=similarity(a, b)$. Then, the basic Omniclust step is:

Omniclust(G)

1) $S \leftarrow \{\}$ (Initialize as empty the set of edges to be preserved.)

2) For each v in V do

a) Let $edges(v)$ be the set of all edges connecting v to other vertices.

b) Let s be the heaviest edge in $edges(v)$

c) $S \leftarrow S \cup \{s\}$

3) E←S (Deletes all edges that were not selected for preservation by any node inspection above. After this step, G will typical be portioned 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 (along with an inter-cluster similarity metric), and the process can be repeated until Omniclust produces just one cluster, which will be the root of a hierarchical clustering based on graph-partitioning in each level.

However, the analysis of results presented here is restricted to clustlets. This choice comes not only from the easy analysis of small-sized clusters, but also as a natural outcome from the relatively small number of features clustered in our experiments.

Finally, in the context of utilization-based clustering, we have chosen to run Omniclust using 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 MOSES-generated 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 MUTIC and MOBRA vectors

Table 1. Examples of MOSES models.

Dataset	Example Models
Aging Brain	or(32052_at NM_006108) or(32052_at NM_006272) or(32052_at not(NM_130463)) or(32052_at not(NM_002576)) or(32052_at not(NM_001217)) or(NM_000518 NM_006108) or(NM_000518 41720_r_at) or(NM_000518 not(1217_g_at)) or(NM_000518 not(NM_001217)) or(NM_000518 not(1558_g_at))
Lymphoma	or(M14328_s_at NM_005566 not(NM_005292)) or(NM_001428 NM_021130 not(NM_005292)) or(NM_001428 NM_005566 not(NM_005292)) or(NM_001428 HG1980-HT2023_at not(NM_005292)) or(NM_002306 HG1980-HT2023_at not(NM_005292)) or(NM_194327 HG1980-HT2023_at not(NM_005292)) or(NM_145902 not(NM_005292) not(NM_002989)) or(NM_145901 not(NM_005292) not(NM_002989)) or(NM_005566 NM_002306 not(NM_005292)) or(NM_005566 not(NM_005292) NM_002629)

will tend to be sparse. Cosine similarity is often used in other machine learning domains involving sparse vectors (such as text classification using word frequency vectors [31]) due to its capacity to deal well with sparseness and produce meaningful similarity relations.

3. RESULTS

MOSES generated 40 "best" models for the Lymphoma dataset, and 88 for aging. For illustration, a few of them are displayed in Table 1, where feature codes correspond to DNA sequences. As one can see, MOSES was able to produce very small Boolean automata for the datasets analyzed. Yet despite their simplicity, MOSES models achieved high out-of-sample accuracies – 94.6% for Lymphoma and 95.3% for Aging Brain.

Model-based vectors were generated from the models using both MUTIC and MOBRA approaches, and those were clustered. Clustering results are evaluated from both quantitative and qualitative standpoints in the following sections.

3.1 Quantitative Analysis

Clustering is a qualitative data analysis method; there are no robust, commonly accepted, objective metrics for comparing different clustering algorithms to each other. [8] 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 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 higher quality than larger ones, which is an undesirable bias.)

By comparing MUTIC and MOBRA to traditional expression-based clustering, according to this cluster quality metric, we found that model-based clustering often produces clearer clusters, with roughly 10 to 100 times greater quality, as shown in the first three rows of Table 2.

This comparison, however, is somewhat unfair to the standard method, because model-based vectors are binary and tending to sparseness, and therefore numerically very different from the non-sparse, real-valued gene expression vectors. In order to detect a potential unfair advantage based on those characteristics, we applied two different binarization 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. This policy may be of special interest since MOSES itself binarizes features by median thresholding.

Using any one of those sparseness policies raises the quality of the expression-based clustering to the same order of magnitude as the utility-based clustering. Still, MOBRA and MUTIC keep a margin of superiority, albeit not quite as dramatic.

3.2 Qualitative Analysis

Next we briefly investigate the qualitative biological significance of the clusters found by our methods. Due to space limitations the analysis is necessarily restricted to a handful of highly salient observations.

Although focused on qualitative biological aspects, the analysis presented below also makes use of a quantitative tool: the GO::TermFinder package [5]. This package computes the p-values of associations between genes and Gene Ontology (GO) categories. Since the GO is a carefully constructed hierarchy of genes based on biological knowledge, such estimates from GO::TermFinder are very helpful in evaluating what biological insights lie in a given cluster. However, they cannot be used to conclude the insignificance of a cluster, only the significance, because of the currently incomplete nature of the GO ontology.

Table 2. Model-based versus expression-based clustering.

Approach	Quality of the Best Cluster	
	Aging Brain	Lymphoma
MUTIC	0.6454	0.5905
MOBRA	0.4375	0.4077
Gene expression	0.0045	0.1995
Gene expression, average binarization	0.1657	0.4071
Gene expression, median binarization	0.2827	0.3859

3.2.1 Lymphoma Dataset

As will be detailed below, MUTIC and MOBRA clustered features with a bias to metabolic processes, especially to "glycolysis". This result has biological support as these pathways are known to be altered in cancers. Comparing the two techniques, MUTIC tends to have a slightly better performance as it generated less clusters and obtained the best statistical result (cluster 6).

3.2.1.1 MOBRA

MOBRA clustering produced 8 clusters, listed in Table 3. Dataset features, originally DNA sequences, are mapped to their corresponding genes. Features that could not be associated to any gene or described are omitted. Since the mapping of sequences to gene is not one-to-one, a given gene may appear in more than one cluster or even more than once in the same cluster. Those observations are also valid for all remaining tables showing clustering results in this paper.

The first 4 clusters did not have much biological information and have higher p values (near or more than 1). Cluster 5 (HSPD1, ALDOA, MIF and GM2A) obtained higher p-values for "cellular lipid metabolic process" (0.31; MIF and GM2A) and "carbohydrate metabolic process" (0.44; ALDOA and GM2A).

Table 3. MOBRA clustering of Lymphoma dataset.

#	Genes
1	CCL21: chemokine (C-C motif) ligand 21
	HMGA1: high mobility group AT-hook 1
2	MT2A : metallothionein 2A
	LGALS3: lectin, galactoside-binding, soluble, 3 (galectin 3)
3	GPR18: G protein-coupled receptor 18
4	PPIA: peptidylprolyl isomerase A (cyclophilin A)
	ENO1: enolase 1, (alpha)
5	MIF: macrophage migration inhibitory factor (glycosylation-inhibiting factor)
	ALDOA: aldolase A, fructose-bisphosphate
	HSPD1: heat shock 60kDa protein 1 (chaperonin)
	GM2A: GM2 ganglioside activator
6	LGALS3: lectin, galactoside-binding, soluble, 3 (galectin 3)
	ENO1: enolase 1, (alpha)
	PKM2 (NM_002654): pyruvate kinase, muscle
7	LDHA : lactate dehydrogenase A
	PKM2 : pyruvate kinase, muscle;
	PGAM1 : phosphoglycerate mutase 1 (brain)
8	PKM2 (NM_182471): pyruvate kinase, muscle
	ALDOA : aldolase A, fructose-bisphosphate
	IFI30 : interferon, gamma-inducible protein 30
	ALDOA : aldolase A, fructose-bisphosphate
	ITGA4 : integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)

All features from cluster 5 are present in "primary metabolic process" with a non significant p-value (1; HSPD1, ALDOA, MIF and GM2A). We know that metabolic processes are affected in cancer and tumors (recently reviewed by [26]). Clusters 6 to 8 reinforce this idea as they all obtained lower and more significant p-values for the "glycolysis" category (0.0041, 2 out of 3 features: ENO1 and PKM1 - cluster6; 2.74e-6, 3 out of 3 features: LDHA, PGAM1 and PKM2 - cluster 7; 0.016, 2 out of 5 features: ALDOA and PKM2 - cluster8). Cluster 8 also have features within "developmental process" (ITGA4, CTSB) and "response to stimulus" (IFI30, CTSB) but these with near one p-values. However, Integrins (ITGA4) are a type of cell-adhesion molecules known to have roles in many types of cancers (reviewed by [4]) and Cathepsin-L (of the family of CTSB) are involved in lymphoid organ regulation [14].

3.2.1.2 MUTIC

MUTIC clustering produced 6 clusters on the Lymphoma dataset. Those are listed in Table 4.

Clusters 1 and 2 do not show much biological information or GO categorization support. But from cluster 3 on, there is a strong support for metabolism, especially "glycolysis", as with the MOBRA results. Clusters 3 and 4 have higher p-values, but with a bias to "metabolic process". Cluster 3 (ALDOA, HSPD1), had a non-significant p-value to "macromolecule metabolic process" and cluster 4 (MIF, GM2A) has a p-value of 0.14 to "cellular lipid metabolic process". Cluster 5 had a p-value of 0.0087 for "glycolysis" with ALDOA and PKM2, and also joined within "developmental process" the features ITGA4 and CTSB, just like MOBRA in its cluster 8.

Cluster 6 grouped 11 features and, from these, 5 have a strong support to 'glycolysis' category with a p-value of 9.51e-08 (LDHA, ALDOA, PGAM1, ENO1, PKM2). This is the best result obtained by both MUTIC and MOBRA. Also, "cell organization and biogenesis" was annotated with low significance to TUBB and ENO1. This category is broadly known to be disrupted in cancers and tumors.

3.2.2 Aging Brain Dataset

In the Aging Brain Dataset, MOBRA and MUTIC mutually corroborated in many GO categorizations. Their best cluster categorization to "negative regulation of apoptosis" has relevance to the dataset. In sum MUTIC and MOBRA had equal performance, with categorizations with more or less the same significance. A more detailed exposition is presented below.

3.2.2.1 MOBRA

MOBRA generated nine clusters on the Aging Brain dataset, as shown in Table 5.

Table 4. MUTIC clustering of Lymphoma dataset.

#	Genes
1	PPIA : peptidylprolyl isomerase A (cyclophilin A)
	ENO1 : enolase 1, (alpha)
2	CCL21 : chemokine (C-C motif) ligand 21
	HMGA1 : high mobility group AT-hook 1
3	HSPD1 : heat shock 60kDa protein 1 (chaperonin)
	ALDOA : aldolase A, fructose-bisphosphate
4	PKM2 : pyruvate kinase, muscle
	GM2A : GM2 ganglioside activator
	MIF : macrophage migration inhibitory factor (glycosylation-inhibiting factor)
5	PKM2 : pyruvate kinase, muscle
	ITGA4 : integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)
	ALDOA : aldolase A, fructose-bisphosphate
6	GPR18 : G protein-coupled receptor 18
	MT2A : metallothionein 2A;
	LGALS3 : lectin, galactoside-binding, soluble, 3 (galectin 3)
	MT2A : metallothionein 2A;
	LDHA : lactate dehydrogenase A
	LGALS3 : lectin, galactoside-binding, soluble, 3 (galectin 3)
	ENO1 : enolase 1, (alpha)
	PKM2 : pyruvate kinase, muscle
	IFI30 : interferon, gamma-inducible protein 30
	PGAM1 : phosphoglycerate mutase 1 (brain)
	ALDOA : aldolase A, fructose-bisphosphate;

Table 5. MOBRA clustering of Aging Brain dataset.

#	Genes
1	VDAC1 : voltage-dependent anion channel 1
	carbonic anhydrase XI (CA11), mRNA.
	ATPase, H+ transporting, lysosomal 13kDa
	HBB : hemoglobin, beta
	glutamate receptor, ionotropic
2	PAK1 : p21/Cdc42/Rac1-activated kinase 1
	CALM1 : calmodulin 1
	DUSP3 : dual specificity phosphatase 3
	spondin 1, extracellular matrix protein (SPON1)
3	PAK1 : p21/Cdc42/Rac1-activated kinase 1
	synapsin II (SYN2), transcript variant IIb, mRNA.
	OGT : O-linked N-acetylglucosamine) transferase
	FADS1 : fatty acid desaturase 1
4	INPP4A : inositol polyphosphate – 4 - phosphatase
	hemoglobin, beta (HBB), mRNA.
	ZNF238 : zinc finger protein 238
	PCMT1 : protein-L-isoaspartate (D-aspartate) O - methyltransferase;
5	50h7 Human retina cDNA randomly primed sublibrary Homo sapiens cDNA, mRNA sequence;
	protein kinase C, zeta (PRKCZ), mRNA.
	Thy-1 cell surface antigen (THY1), mRNA.
	Thy-1 co-transcribed (LOC94105), mRNA.
	annexin A4 (ANXA4), mRNA.
6	HBB : hemoglobin, beta;CD113t-C, HBD
	RAB6A : RAB6A, member RAS oncogene family
7	VAMP1 : vesicle-associated membrane protein 1
	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B) (ELAVL2), mRNA.
8	p21/Cdc42/Rac1-activated kinase 1
	syndecan 2 (heparan sulfate proteoglycan 1, cell surface-associated, fibroglycan) (SDC2), mRNA.
9	MAP2 : microtubule-associated protein 2
	HLF : hepatic leukemia factor
	S100 calcium binding protein, beta (neural)
	adaptor-related protein complex 1, sigma 1 subunit (AP1S1), transcript variant 2, mRNA.

The GO::TermFinder package was used to evaluate MOBRA clusters with GO categorization within "biological process". The most significant cluster obtained by MOBRA was cluster 5, with a p-value of 0.00019. It was categorized as "negative regulation of apoptosis" based on 3 out of 4 features (ANXA4, THY1, PRKCZ). By definition, apoptosis is an event linked to aging by many aspects. A deregulated apoptosis system can lead to cancers or neurodegenerative diseases, common in aging (reviewed by [1]).

Four other clusters obtained lower p-values with little significance and biological importance yet to be explored. Cluster #1 obtained a p-value of 0.07 for "ion transport" features (ATP6V1G2, VDAC1, GRIN2A). Cluster #2, with the features DUSP3 and PAK1 scored a p-value of 0.64618 for "phosphate metabolic process". Cluster #6 for "transport" with the features HBB and RAB6A. Cluster #9 with a p-value of 0.11 for 2 out its 5 features (S100B and MAP2) to "regulation of protein metabolic process".

Clusters #3, 4, 7 and 8 scored low significance (p-value equal or over 1.0) according to GO::TermFinder and will not be discussed. They may of course have biological significance not revealed by the current state of the GO database.

3.2.2.2 MUTIC

MUTIC applied to the Aging brain dataset produced ten clusters, shown in Table 6.

The most significant MUTIC cluster supported the best MOBRA cluster result: cluster #2 was also categorized to "negative regulation of apoptosis" with a significant p-value of 0.007 (little less significant than the correspoding MOBRA value), and with 2 same features (ANXA4, THY1). The other feature was left to another cluster with no significant categorization.

MUTIC also generated a more significant cluster, #10, with a p-value of 0.02. This cluster has a total of 6 features, 4 of them categorized to "phosphate metabolic process" (PAK1, DUSP3, THY1, PRKCB1). The features PAK1 and DUSP3 was also part of a cluster with the same categorization from MOBRA, but with less significant p-value. The biological importance of the GO category corresponding to these two clusters was already discussed in the MOBRA section.

Three other clusters had less significant categories. Cluster #8, p-value of 0.38, for "transport" also joined the HBB and RAB6A features like MOBRA, but another feature VDAC1 was added to this category. Cluster #4 for "regulation of biological process" with the features "ELAVL2, SYN2, MAP2" and a p-value of 0.34. Cluster #9 for "biosynthetic process", with a p-value of 0.39 and 3 out of 5 features in this category (OGT, ATP6V1G2, HBB). The other clusters had no significance according to GO::TermFinder, and will not be discussed.

4. Conclusions

We have presented a novel analytical method, model-based clustering, and illustrated its behavior by discussing the results of its application to two test datasets, and using two different model-based mappings. In both cases the method has shown itself able to identify interesting inter-gene, inter-process and condition-related interactions.

Like standard expression-based clustering, this 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.

Table 6. MUTIC clustering of Aging Brain dataset.

Quality	Genes
0.6454	PAK1 : p21/Cdc42/Rac1 - activated kinase 1
	SDC2 : syndecan 2
0.6268	THY1 : Thy - 1 cell surface antigen
	ANXA4 : annexin A4
0.5375	ANK2 : ankyrin 2, neuronal
	RHOBTB3 : Rho-related BTB domain containing 3
	GRIN2A : glutamate receptor, ionotropic, N - methyl D - aspartate 2A
0.5335	MAP2 : microtubule - associated protein 2
	SYN2 : synapsin II;SYNII, SYNIIa, SYNIIb
	ELAVL2 : ELAV - like 2 (Hu antigen B)
0.5306	SPON1 : spondin 1, extracellular matrix protein
	PCMT1:protein-L-isoaspartate O- methyltransferase
	FADS1 : fatty acid desaturase 1
	HLF : hepatic leukemia factor
0.5105	AP1S1:adaptor-related protein complex 1,sigma 1
	S100B : S100 calcium binding protein B;
0.4912	VAMP1 : vesicle - associated membrane protein 1
	50h7 Human retina cDNA randomly primed
	PRKCZ : protein kinase C, zeta
0.4859	RAB6A : RAB6A, member RAS oncogene family
	HBB : hemoglobin, beta
	CA11 : carbonic anhydrase XI
	HBB : hemoglobin, beta
	VDAC1 : voltage - dependent anion channel 1
0.4634	ATP6V1G2 : ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G2
	INPP4A : inositol polyphosphate - 4 - phosphatase
	ATP6V1G2 : ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G2
	ZNF238 : zinc finger protein 238
	HBB : hemoglobin, beta
	OGT : O - linked N - acetylglucosamine transferase
0.4583	THY1 : Thy - 1 cell surface antigen
	KLHDC3 : kelch domain containing 3
	DUSP3 : dual specificity phosphatase 3
	PRKCB1 : protein kinase C, beta 1
	PAK1 : p21/Cdc42/Rac1 - activated kinase 1
	CALM1 : calmodulin 1
	PAK1 : p21/Cdc42/Rac1 - activated kinase 1

However, there is also an objective component to the advantage of the present approach over traditional clustering, in that there are some types of interrelationships that utilization-based clustering can capture, which traditional expression-based clustering is mathematically unable to.

A note should be made on the use of MOSES in these experiments, rather than some other supervised categorization method. A requirement for model-based clustering is a categorization method that learns a variety of different models for a single problem, in which the different models have differences in feature utilization that are easy to detect. In [11] we have reported some results obtained using standard Genetic Programming (GP) together with the MUTIC methodology; however, for the further explorations given here with MUTIC and MOBRA, we chose to switch from GP to MOSES. The reason for this switch was the improved compactness of the MOSES models. GP models often suffer from "bloat," a phenomenon which often leads to the incorporation into a model of features that aren't really essential to the model's accuracy, but are difficult to remove via simplistic pruning methods. Being smaller, MOSES models introduce less noise into the model-based clustering process. A systematic comparison of model-based clustering using GP and MOSES goes beyond the scope of this paper, but qualitatively speaking we have observed that the MOSES-based model-based clustering results tend to have higher clustering quality and more biological relevance.

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 model-based clustering 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 may be meaningfully treated as categorical and that displays complex inter-feature interactions.

5. REFERENCES

[1] Alenzi F. Q. Apoptosis and diseases: regulation and clinical relevance. *Saudi Med J,* 26, 11 (Nov 2005), 1679-90.

[2] Bar-Joseph Z., Demaine E.D., Gifford D.K., Srebro N., Hamel A.M., Jaakkola T.S. K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics* 19: 1070-1078, 2003.

[3] Ben-Dor A., Shamir R., Yakhini Z. Clustering gene expression patterns. *J Comput Biol* 6: 281-297, 1999.

[4] Bogenrieder T., Herlyn M. Axis of evil: molecular mechanisms of cancer metastasis. *Oncogene*, 22, 42 (Sep 2003), 6524-36.

[5] Boyle E.I., Weng S., Gollub J., Jin H., Botstein D., Cherry J. M., Sherlock G. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*. 20, 18 (Dec 2004), 3710-5.

[6] Brown M.P., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M., Jr., Haussler D.. Knowledge-based analysis of microarray gene expression data by using

support vector machines. *Proc Natl Acad Sci U S A* 97: 262-267, 2000.

[7] Cho J.H., Lee D., Park J.H., Lee I.B. Gene selection and classification from microarray data using kernel machine. *FEBS Lett* 571: 93-98, 2004.

[8] Dopazo J., Azuaje F. Data analysis and visualization in genomics and proteomics. John Wiley, Chichester, West Sussex; Hoboken, NJ, 2005.

[9] Dudoit S., Fridlyand J., Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 97: 77-87, 2002.

[10] Eisen M.B., Spellman P.T., Brown P.O., Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95: 14863-14868, 1998.

[11] Goertzel B., Pennachin C., de Souza Coelho L., Mudado M. Identifying Complex Biological Interactions based on Categorical Gene Expression Data. In Gary G. Yen and Lipo Wang and Piero Bonissone and Simon M. Lucas editors, Proceedings of the 2006 IEEE Congress on Evolutionary Computation, pages 5583-5590, Vancouver, 2006. details

[12] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A.., Bloomfield C.D., Lander E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537, 1999.

[13] Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning* 46: 389-422, 2002.

[14] Lombardi G, Burzyn D, Mundinano J, Berguer P, Bekinschtein P, Costa H, Castillo LF, Goldman A, Meiss R, Piazzon I, Nepomnaschy I. Cathepsin-L influences the expression of extracellular matrix in lymphoid organs and plays a role in the regulation of thymic output and of peripheral T cell number. *J Immunol*, 174, 11 (Jun 2005), 7022-32.

[15] Looks, M. Competent Program Evolution. PhD thesis, Washington University in St. Louis, 2006.

[16] Lu T., Pan Y., Kao S.Y., Li C., Kohane I., Chan J., Yankner B.A.. Gene regulation and DNA damage in the Aging human brain. *Nature* 429: 883-891, 2004.

[17] Markovetz F.. A bibliography on learning causal networks of gene interactions. 2004

[18] Markovetz F., Spang R. Reconstructing gene regulation networks from passive observations and active interventions. *7th Ann Intl Conf Res Comput Molec Biol (RECOMB),* 2003.

[19] Mattson M.P. Neuronal life-and-death signaling, apoptosis, and neurodegenerative disorders. *Antioxid Redox Signal*. 8, 11-12 (Nov-Dec 2006), 1997-2006.

[20] Nachman I., Regev A., Friedman N.. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* 20 Suppl 1: I248-I256, 2004

[21] Neiman P.E., Ruddell A., Jasoni C., Loring G., Thomas S.J., Brandvold K.A., Lee R., Burnside J., Delrow J. 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.

[22] Segal E., Koller. D. Probabilistic hierarchical clustering for biological data. *6th Ann Intl Conf Res Comput Molec Biol (RECOMB),* 2002.

[23] Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D. and Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9: 3273-3297, 1998.

[24] Sharan R., Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* 8: 307-316, 2000.

[25] Sharan R., Elkon R., Shamir R. Cluster analysis and its applications to gene expression data. *Ernst Schering workshop on Bioinformatics and Genome Analysis.* Springer Verlag, 2001.

[26] Shaw R. J. Glucose metabolism and cancer. *Curr Opin Cell Biol*, 18, 6 (Dec 2006), 598-608.

[27] Shipp M. A., Ross K. N., Tamayo P., Weng A. P., Kutok J. L.. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine.* 2002.

[28] Sohler F., Hanisch D., Zimmer R. New methods for joint analysis of biological networks and expression data. *Bioinformatics* 20: 1517-1521, 2004.

[29] Statnikov, A., Aliferis, C. F., Tsamardinos, I., D. Hardins and S. Levy. (2005) 'A comprehensive evaluation of multicategory classification methods for microarray gene expression diagnosis', Bioinformatics. 21, 5, 631-643.

[30] Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub T.R.. "nterpreting 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.

[31] Van Rijsbergen C.J.. Information retrieval. Butterworths, London; Boston, 1979.

[32] Vert J.P., Kanehisa M. Extracting active pathways from gene expression data. *Bioinformatics* 19 Suppl 2: II238-II244, 2003.