RESEARCH ARTICLE OPEN ACCESS

Cluster Analysis for Gene Expression Data: A Survey

Bhupesh Deka1, Sitanath Biswas2, Narendra Kumar Rout3
[1,2]*Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar*
[3] *Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar*

Abstract
DNAmicroarraytechnologyhasnowmadeitpossibletosimultaneouslymonitortheexpres- sion levels of thousands of genes during important biological processes and across collections of related samples. Elucidating the patterns hidden in gene expression data offers a tremen- dous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. A first step toward addressing this challenge is the use of clustering techniques, which is essential in the data mining process store vealnatural structures and identify interesting patterns in the underlying data.

Cluster analysis seeks to partition a given data set into groups based on specified features sothatthedatapoints with in a group are more similar to each other than the points in different groups. A very rich literature on cluster analysis has developed over the past three decades. Many conventional clustering algorithms have been adapted or directly applied to gene expres- siondata, and alsonew algorithms have recently been proposed specifically aiming at geneex- pression data. These clustering algorithms have been proven useful for identifyingbiologically relevant groups of genes and samples.

In this paper, we first briefly introduce the concepts of microarray technology and discuss the basic elements of clustering on gene expression data. Inparticular,wedivideclusteranalysis for gene expression data into three categories. Then we present specific challenges pertinent to each clustering category and introduce several representative approaches. We also discuss the problem of cluster validation in three aspects and review various methods to assess the quality and reliability of clustering results. Finally, we conclude this paper and suggest the promising trends in this field.

I. INTRODUCTION

Introduction to MicroarrayTechnology

Measuring mRNAlevels

Comparedwiththetraditionalapproachtogenomicresearch,which has focuse donth eloca lexami- nation and collection of data on singlegenes,micro array technologies have now made it possible to monitor the expression levels for tens of thousands of genes in parallel. The two major types ofmi- croarray experiments are the cDNA microarray [54] and oligonucleotide arrays (abbreviated oligo chip)[44]. Despitedifferencesinthedetailsoftheirexperimentprotocols,bothtypesofexperiments involve three common basic procedures[67]:

Chip manufacture: A microarray is a small chip (made of chemically coated glass, nylonmembrane or silicon), onto which tens of thousands of DNA molecules (probes) are attached in fixed grids. Each grid cell relates to a DNA sequence.

Target preparation, labeling and hybridization: Typically, two mRNA samples (a test sample and a control sample) are reverse-transcribed into cDNA (targets), labeled using either fluo- rescent dyesor radio active isotopics, and then hybridized with the probes on the surface of the chip.

The scanning process: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

Generally, both cDNA microarray and oligo chip experiments measure the expression level for each DNA sequence by the ratio of signal intensity between the test sample and the controlsample, therefore, data sets resulting from both methods share the same biological semantics. In this paper, unless sexplicitly stated, we will refertoboththecDNA micro array and the oligochip as micro array technology and term the measurements collected via both methods as gene expressiondata.

Pre-processing of gene expressiondata

A micro array experiment typically assesses a large number of DNA sequences (genes, cDNAclones, or expressed sequence tags [ESTs]) under multiple conditions. These conditions may be a time- series during a biological process (e.g., the yeast cell cycle) or a collection of different tissue sam- ples (e.g., normal versus cancerous tissues). In this paper, we will focus on the cluster analysis of gene expression data withou t making adistinction among DNA sequences, which will uniformly be called “genes”. Similarly, we will uniformly refer to all kinds of experimental conditions as “sam- ples” if no confusion will be caused. A gene expression data set from a microarray experiment can be represented by a real-valued **expression matrix** (Fig- ure 1(a)), where the rows () form the expression patterns of genes, the columns () represent the expression profiles of samples, and each cell is the measured expression level of gene in sample . Figure 1 (b) includes some notation that will be used in the followingsections.

The original gene expression matrix obtained from a scanning process contains noise, missing values, and systematic variations arising from the experimental procedure. Data pre-processing is indispensable before any cluster analysis can be performed. Some problems of data pre-processing have themselves become interesting research topics. Those questions are beyond the scope of this survey; an examination of the problem of missing value estimation appearsin[69], and the problem of data normalization is addressed in [32, 55]. Furthermore, many clustering approaches apply one or more of the following pre-processing procedures: filtering out genes with expression levels which do not change significantly across samples; performing a logarithmic transformation ofeach expression level; or standardizing each row of the gene expression matrix with a mean of zero and a variance of one. In the following discussion of clustering algorithms, we will set aside the details of pre-processing procedures and assume that the input data set has already been properly pre-processed.

Applications of clustering gene expressiondata

Clusteringtechniqueshaveproventobehelpfultounderstandgenefunction,generegulation,cellularprocesses,andsubtypesofcells.Geneswithsimilarexpressionpatterns(coexpressedgenes)canbeclusteredtogetherwithsimilarcellularfunctions.Thisapproachmayfurtherunderstandingofthefunctionsofmanygenesforwhichinformationhasnotbeenpreviouslyavailable[66,20].Furthermore,coexpressedgenesinthesameclusterarelikelytobeinvolvedinthesamecellularprocesses,andastrongcorrelationofexpressionpatternsbetweenthosegenesindicatesco-regulation.Search- ing for common DNA sequences at the promoter regions of genes within the same cluster allows regulatorymotifsspecifictoeachgeneclustertobeidentifiedandcis-regulatoryelementstobepro- posed [9, 66]. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network [16]. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches [2,24].

Introduction to ClusteringTechniques

In this subsection, we will first introduce the concepts of clusters and clustering. We will then divide the clustering tasks for gene expression data into three categories according to different clustering purposes. Finally, we will discuss the issue of proximity measure indetail.

Clusters andclustering

Clustering is the process of grouping data objects into a set of disjoint classes, called clusters, so that objects within a class have high similarity to each other, while objects in separate classes are more dissimilar. Clustering is an example of unsupervised classification. “Classification” refers to a procedure that assigns data objects to a set of classes. “Unsupervised” means that clustering does not rely on predefined classes and training examples while classifying the data objects. Thus, clustering is distinguished from pattern recognition or the areas of statistics known as discriminant analysis and decision analysis, which seek to find rules for classifying objects from a given set of pre-classified objects.

Categories of gene expression dataclustering

Currently, a typical microarray experiment contains to genes, and this number is expected to reach to the order ofHowever, the number of samples involved in a micro array experimentis generally less than **One of the characteristics of gene expression data is that it ismeaningful to cluster both genes and samples.** On one hand, co-expressed genes can be grouped in clusters based on their expression patterns [7, 20]. In such **gene-based clustering**, the genes are treated as the objects, while the samples are the features. On the other hand, the samples can be partitioned into homogeneous groups. Each group may correspond to some particular macroscopic phenotype, such as clinical syndromes or cancer types[24]. Such **sample-based clustering** regards the samples as the objects and the genes as the features. The distinction of gene-based clustering and sample-based clustering is based on different characteristics of clustering tasks for gene expression data. Some clustering algorithms, such as K-means and hierarchical approaches, can be used both to group genes and to partition samples. We will introduce those algorithms as gene-based clustering approaches, and will discuss how to apply the mas sample-based clustering in subsection 2.2.1.

Both the gene-based and sample-based clustering approaches search exclusive and exhaustive partitions of objects that share the same feature space. However, current thinking in molecular biology holds that only a small subset of genes participate in any cellular process of interest and that a cellular process takes place only in a subset of the samples. This belief calls for the **subspace clustering** to capture clusters formed by a subset of genes across a subset of samples. For subspace clustering algorithms, genes and samples are treated symmetrically, so that either genes or samples canberegardedasobjectsorfeatures.Furthermore,clustersgeneratedthroughsuchalgorithmsmay have different featurespaces.

While a gene expression matrix can be analyzed from different angles, the gene-based, sample-basedclusteringandsubspaceclusteringanalysisfaceverydifferentchallenges.Thus,wemayhave to adopt very different computational strategies in the three situations. The details of thechallenges and the representative clustering techniques pertinent to each clustering category will be discussed in Section2.

Proximity measurement for gene expressiondata

Proximity measurement measures the similarity (or distance) between two data objects. Gene expression data objects, no matter genes or samples, can be formalized as numerical vectors , where is the value of the th feature for the th data object and is the number of features. The proximity between two objects and is measured by a proximity function of corresponding vectorsand.

Euclideandistanceisoneofthemostcommonly-usedmethodstomeasurethedistancebetween two data objects. The distance between objects and in dimensional space is definedas:

However, for gene expression data, the overall shapes of gene expression patterns (or profiles)areofgreaterinterestthantheindividualmagnitudesofeachfeature.Euclideandistancedoesnotscore well for shifting or scaled patterns (or profiles) [71]. To address this problem, each object vector is standardized with zero mean and variance one before calculating the distance [66, 59,56].

AnalternatemeasureisPearson'scorrelationcoefficient,whichmeasuresthesimilaritybetween the shapes of two expression patterns (profiles). Given two data objects and, Pearson'scorrelation coefficient is defined aswhere and are the means for and , respectively. Pearson's correlation coefficient views each object as a random variable with observations and measures thesimilaritybetweentwoobjectsbycalculatingthelinearrelationshipbetweenthedistributionsofthetwocorresponding randomvariables.

Pearson'scorrelationcoefficientiswidelyusedandhas proveneffectiveasasimilaritymeasure for gene expression data [36, 64, 65, 74]. However, empirical study has shown that it is not robust with respect to outliers [30], thus potentially yielding false positives which assign a high similarity score to a pair of dissimilar patterns. If two patterns have a common peak or valley at a single feature,thecorrelationwillbedominatedbythisfeature, althoughthepatternsattheremainingfea- tures may be completely dissimilar. This observation evoked an improved measure calledJackknife correlation [19, 30], defined as , whereis the Pearson's correlation coefficient of data objects and with the lth feature deleted. Use of theJackknifecorrelationavoidsthe"dominanceeffect" ofsingleoutliers.MoregeneralversionsofJackknifecorrelationthatarerobusttomorethanoneoutliercansimilarlybederived.However,the generalized Jackknife correlation, which would involve the enumeration of different combinations of features to be deleted, would be computationally costly and is rarelyused.

Another drawback of Pearson's correlation coefficient is that it assumes an approximate Gaussian distribution of the points and may not be robust for non-Gaussian distributions [14, 16]. To addressthis,theSpearman'srankordercorrelationcoefficienthasbeensuggestedasthesimilarity measure. The ranking correlation is derived by replacing the numerical expression level with its rank among all conditions. For example, if is the third highest value among,where Spearman's correlation coefficient does not require the assumption ofGaussian

distribution and is more robust against outliers than Pearson's correlation coefficient. However, as a consequence of ranking, a significant amount of information present in the data is lost. Our ex- perimental results indicate that, on average, Spearman's rank-order correlation coefficient doesnotperform as well as Pearson's correlation coefficient.

Almost all of the clustering algorithms mentioned in this survey use either Euclidean distance or Pearson's correlation coefficient as the proximity measure. When Euclidean distance is selectedas proximity measure, thestandardizationprocess is usually applied,where is the th feature of object , while and are the mean and standard deviationof , respectively.Supposeandarethestandardized"objects"of and . ThenwecanproveThese two equations disclose the consistency between Pearson's correlation coefficient and Eu- clidean distance after data standardization; i.e., if a pair of data objects , has a highercorrelation than pair then pair , has asmaller distance than pair. Thus, we canexpect theeffectivenessofaclusteringalgorithmtobeequivalentwhetherEuclideandistanceorPearson's correlation coefficient is chosen as the proximitymeasure.

II. CLUSTERING ALGORITHMS

As we mentioned in Section 1.2.2, gene expression matrix can be analyzed in two ways. For gene- based clustering, genes are treated as data objects, while samples are considered as features. Conversely,forsamplebasedclustering,samplesserve asdataobjectstobeclustered,whilegenesplay the role of features. The third category of cluster analysis applied to gene expression data, which is subspace clustering, treats genes and samples symmetrically such that either genes or samples canberegardedasobjectsorfeatures.Genebased,sample-basedandsubspaceclusteringfacevery different challenges, and different computational strategies are adopted for each situation. In this section,wewillintroducethegenebasedclustering,samplebasedclustering,andsubspaceclusteringtechniques,respectively.

Gene-based Clustering

In this section, we will discuss the problem of clustering genes based on their expression patterns. The purpose of gene-based clustering is to group together co-expressed genes which indicate co-function and co-regulation. We will first present the challenges of gene-based clusteringandthenreviewaseriesofclusteringalgorithmswhichhavebeenappliedtogroupgenes.Foreachclustering algorithm, we will first introduce the basic idea of the clustering process, and then highlight some features of thealgorithm.

K-means

The K-means algorithm [46] is a typical partition-based clustering method.Given a pre-specifiednumber **!** ,thealgorithmpartitionsthedataset into **!** disjointsubsetswhichoptimizethefollowing objective function:

Here, is a data object in cluster "and is the centroid (mean of objects) of ". Thus, theobjectivefunctiontriestominimizethesumofthessquareddistancesofobjectsfromtheircluster centers.

The K-means algorithm is simple and fast. The time complexity of K-means is , where is the number of iterations and is the number of clusters. Our empirical study has shown thattheKmeansalgorithmtypicallyconvergesinasmallnumberofiterations.However,italsohasseveraldrawbacksasagenebasedclusteringalgorithm.First,thenumberofgeneclustersinagene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of and compare the clustering results. For a large gene expression data set which contains thousands of genes, this extensive parameter fine-tuning process may not be practical. Second, gene expression data typically contain a huge amount of noise; however, the K-means algorithm forces each gene into a cluster, which maycause the algorithm to be sensitive to noise [59,57].

Recently, several new clustering algorithms [51, 31, 59] have been proposed to overcome the drawbacks of the K-means algorithm. These algorithms typically use some global parameters to control the quality of resulting clusters (e.g., the maximal radius of a cluster and/or the minimal distance between clusters). Clustering is the process of extracting all of the qualified clusters from the data set. In this way, the number of clusters can be automatically determined and thosedataobjectswhichdonotbelongtoanyqualifiedclustersareregardedasoutliers.However,thequalities ofclustersingeneexpressiondatasetsmayvarywidely. Thus,itisoftenadifficultproblemto choose the appropriate globally-constraining parameters.

Self-organizing map

The Self-Organizing Map (SOM) was developed by Kohonen [39], on the basis of a single layered neural network. The data objects are presented at the input, and the output neurons are organized with a simple neighborhood structure such as a two dimensional# grid. Each neuron of the neural network is associated with a reference vector, and each data point is "mapped" to the neuron with the "closest" reference vector. In the process of running the algorithm, each data object actsasatrainingsamplewhichdirectsthemovementofthereferencevectorstowardsthedenserareasofthe inputvectorspace,sothatthosereferencevectorsaretrainedtofitthedistributionsoftheinputdata set. When the training is complete, clusters are identified by mapping all data points to the output neurons.

One of the remarkable features of SOM is that it generates an intuitively-appealing map of a high-dimensional data set in $ or $ space and places similar clusters near each other. Theneuron trainingprocessofSOMprovidesarelativelymorerobustapproachthanK-meanstotheclustering of highly noisy data [62, 29]. However, SOM requires users to input the number of clusters and the grid structure of the neuron map. These two parameters are preserved through the training process; hence, improperly-specified parameters will prevent the recovering of the natural cluster structure. Furthermore, if the data set is abundant with irrelevant data points, such as genes with invariant patterns, SOM will produce an output in which this type of data will populate the vast majority of clusters [29]. In this case, SOM is not effective because most of the interesting patterns may be merged into only one or two clusters and cannot beidentified.

Hierarchicalclustering

Incontrasttopartitionbasedclustering,which attemptstodirectlydecomposethedatasetintoaset of disjoint clusters, hierarchical clustering generates a hierarchical series of nested clusters which can be graphically represented by a tree, called dendrogram. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the dendrogram at some level, we can obtain a specified number of clusters. Byreordering theobjectssuchthatthebranchesofthecorrespondingdendrogramdonotcross,thedatasetcanbe arranged with similar objects placedtogether.

Hierarchical clustering algorithms can be further divided into agglomerative approaches and divisiveapproachesbasedonhowthehierarchicaldendrogramisformed.Agglomerativealgorithms (bottom-up approach) initially regard each data object as an individual cluster, and at each step, mergetheclosestpairofclustersuntilallthegroupsaremergedintoonecluster.Divisivealgorithms (top-down approach) starts with one cluster containing all the data objects, and at each step split a cluster until only singleton clusters of individual objects remain. For agglomerative approaches, different measures of cluster proximity, such as single link, complete link~~andm~~inimumvariance[18,38],derivevariousmergestrategies.Fordivisiveapproaches,theessentialproblemistodecide how to split clusters at each step. Some are based on heuristic methods such as the deterministic annealing algorithm [3], while many others are based on the graph theoretical methods which we will discuss later.

Eisen et al. [20] applied an agglomerative algorithm called UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and adopted a method to graphically represent the clustered data set. In this method, each cell of the gene expression matrix is colored on the basis of the measured fluorescence ratio, and the rows of the matrix are re-ordered based on the hierarchical dendrogram structure and a consistent node-ordering rule. After clustering, the original gene expression matrix isrepresentedbyacoloredtable(aclusterimage)wherelargecontiguouspatchesofcolorrepresent groups of genes that share similar expression patterns over multipleconditions.

Alon et al. [3] split the genes through a divisive approach, called the deterministic-annealing algorithm (DAA) [53, 52]. First, two initial cluster centroids '', , were randomly defined. The expression pattern of gene was represented by a vector, and the probability ofgene belonging to cluster was assigned according to a two-component Gaussian model:
% &''' % &''. The cluster centroids were recalculated by ''
' . An iterative process (the EM algorithm) was then applied to solve
and ''(the details of the EM algorithm will be discussed later). For &, there was only one cluster, ''''. When &was increased in small steps until a threshold was reached, two distinct, converged centroids emerged. The whole data set was recursively split until each clustercontained
only one gene.

Hierarchical clustering not only groups together genes with similar expression pattern but alsoprovidesanaturalwaytographicallyrepresentthedataset.Thegraphicrepresentationallowsusers athoroughinspectionofthewholedatasetandobtainaninitialimpressionofthedistributionofdata. Eisen’s method is much favored by many biologists and has become the most widely-used tool in geneexpressiondataanalysis[20,3,2,33,50].However, theconventionalagglomerativeapproach suffersfromalackofrobustness[62],i.e.,asmallperturbationofthedatasetmaygreatlychangethe structure of the hierarchical dendrogram. Another drawback of the hierarchical approach is its high computationalcomplexity.Toconstructa“complete”dendrogam(whereeachleafnodecorresponds toonedataobject,andtherootnodecorrespondstothewholedataset),theclusteringprocessshould takemerging (or splitting) steps. The time complexity for a typical agglomerative hierarchical algorithm is [34]. Furthermore, for both agglomerative and divisive approaches, the “greedy” nature of hierarchical clustering prevents the refinement of the previous clustering. If a “bad” decision is made in the initial steps, it can never be corrected in the followingsteps.

Graph-theoreticalapproaches

Givenadataset❨,wecanconstructaproximitymatrix,where%)*,andaweightedgraph+,calledaproximitygraph,whereeachdatapointcorrespondstoavertex. Forsomeclusteringmethods,eachpairofobjectsisconnectedbyanedgewithweightassigned accordingtotheproximityvaluebetweentheobjects[56,73].Forothermethods,proximityis mappedonlytoeitherorontheb asisofsomethreshold,andedgesonlyexistbetweenobjects and , where equals [7, 26]. Graph-theoretical clustering techniques are explicitly presented intermsofagraph,thusconvertingtheproblemofclusteringadatasetintosuchgraphtheoretical problems as finding minimum cut or maximal cliques in the proximity graph.

CLICK. CLICK (CLuster Identification via Connectivity Kernels) [56] seeks to identify highly connected components in the proximity graph as clusters. CLICK makes the probabilistic assumption that after standardization, pair-wise similarity values between elements (no matter they are in the same cluster or not) are normally distributed. Under this assumption, the weight ¸ of an edgeis defined as the probability that vertices andare in the same cluster. The clustering process of CLICK iteratively finds the minimum cut in the proximity graph and recursively splits the data set into a set of connected components from the minimum cut. CLICK also takes two post-pruning steps to refine the cluster results. The adoption step handles the remainingsingletonsandupdatesthecurrentclusters,whilethemergingstepiterativelymergestwoclusterswithsimilarityexceeding a predefinedthreshold.

In [56], the authors compared the clustering results of CLICK on two public gene expression data sets with those of GENECLUSTER [62] (a SOM approach) and Eisen's hierarchical approach [20],respectively.Inbothcases,clustersobtainedbyCLICKdemonstratedbetterqualityintermsofhomogeneity and separation (these two concepts will be discussed in Section 3). However, CLICK has little guarantee of not going astray and generating highly unbalanced partitions, e.g., apartition that only separates a few outliers from the remaining data objects. Furthermore, in gene expression data, two clusters of co-expressed genes, "and ", may be highly intersected with each other. In such situations, "and "are not likely to be split by CLICK, but would be reported as one highly connected component.

CAST. Ben-Dor et al. [7] introduced the idea of a corrupted clique graph data model. The input data set is assumed to come from the underlying cluster structure by "contamination" with random errors caused by the complex process of gene expression measurement. Specifically, it is assumed that the true clusters of the data points can be represented by a clique graph , which is a disjoint union of complete sub-graphs with each clique corresponding to a cluster. The similarity graphis derived from by flipping each edge/non-edge with probability -. Therefore, clustering a dataset is equivalent to identifying the original clique graph from the corrupted version with as few flips (errors) aspossible.

In [7], Ben-Dor et al. presented both a theoretical algorithm and a practical heuristic called CAST(ClusterAffinitySearchTechnique).CASTtakesasinputareal,symmetric,n-by-nsimilarity matrix (and an affinity threshold). The algorithm searches the clusters one at a time. The currently searched cluster is denoted by ". Each element % has an affinity value % with respect to "as % % * . An element % has a high affinity value if itsatisfies %) "; otherwise, % has a low affinity value. CAST alternates between adding high-affinity elements to the current cluster, and removing low-affinity elements from it. When the process stabilizes, "is considered a complete cluster, and this process continues with each new cluster until all elements have been assigned to a cluster.

The affinity threshold) of the CAST algorithm is actually the average of pairwise similarities within a cluster. CAST specifies the desired cluster quality through) and applies a heuristicsearch- ing process to identify qualified clusters one at a time. Therefore, CAST does not depend on a user-defined number of clusters and deals with outliers effectively. Nevertheless, CAST has the usual difficulty of determining a "good" value for the global parameter).

Model-basedclustering

Modelbasedclusteringapproaches[21,76,23,45]provideastatisticalframeworktomodeltheclusterstructureofgeneexpressiondata.Thedatasetisassumedtocomefromafinitemixtureofunderlying probability distributions, with each component corresponding to a different cluster. The goal is to estimate the parameters . andthat maximize the likelihood where is the number of dataobjects,is the number of components, % is a data object (i.e., a gene expression pattern), % . is the densityfunctionof%ofcomponent"withsomeunknownsetofparameters.(modelparameters), and / (hidden parameters) represents the probability that % belongs to ". Usually, theparameters and are estimated by the EM algorithm. The EM algorithm iterates between Expectation (E) steps and Maximization (M) steps. In the E step, hidden parameters are conditionally estimated from the data with the current estimated . In the M step, model parameters are estimated so as to maximize the likelihood of complete data given the estimated hidden parameters. When the EMalgorithmconverges,eachdataobjectisassignedtothecomponent(cluster)withthemaximum conditionalprobability.Animportantadvantageofmodelbasedapproachesisthattheyprovideanestimatedprobabilitythat data object will belong to cluster. As we will discuss in Subsection 2.1.1, gene expression data are typically "highly-connected"; there may be instances in which a single gene has a highcorrelationwithtwodifferentclusters.Thus,theprobabilisticfeatureofmodel-basedclusteringis particularly suitable for gene expression data. However, model-based clustering relies on the assumption that the data set fits a specific distribution. This may not be true in many cases.

Themodelingofgeneexpressiondatasets,inparticular, isanongoingeffortbymanyresearchers,and,tothebest ofourknowledge,thereiscurrentlynowellestablished modeltorepresentgeneexpression data. Yeung et al. [76] studied several kinds of commonly-used data transformationsandassessedthedegreetowhichthreegeneexpressiondatasetsfitthemultivariantGaussianmodelassumption. Therawvaluesfromallthreedatasets fittheGaussianmodelpoorlyandthereisnouniformrule to indicate which transformation would best improve thisfit.

A density-based hierarchical approach:DHC

In[36],theauthorsproposedanewclusteringalgorithm,DHC(adensitybased,hierarchicalclustering method), to identify the co-expressed gene groups from gene expression data. DHC is developed based on the notions of “density” and “attraction” of data objects. The basic idea is to consider a cluster as a high-dimensional dense area, where data objects are “attracted” with each other. At the “core”partofthedensearea,objectsarecrowdedclosely witheachother,andthushavehighdensity.

Objects at the peripheral area of the cluster are relatively sparsely distributed, and are “attracted” to the “core” part of the dense area.

Oncethe“density”and“attraction”ofdataobjectsaredefined,DHCorganizestheclusterstruc- ture of the data set in two-level hierarchical structures. At the first level, an attraction tree is con- structed to represent the relationship between the data objects in the dense area. Each node on the attraction tree corresponds to a data object, and the parent of each node is its attractor. The only exception is the data object which has the highest density in the data set. This data object becomes the root of the attraction tree. However, the structure of the attraction tree would be hard to inter- pret when the data set becomes large and the data structure becomes complicated. To address this problem, at the second structure level, DHC summarizes the cluster structure of the attraction tree into a density tree. Each node of the density tree represents a dense area. Initially, the whole data set is considered as a single dense area and is represented by the root node of the density tree. This dense area is then split into several sub-dense areas based on some criteria, where each sub-dense areaisrepresentedbyachildnodeoftherootnode.These sub-denseareasarefurthersplit,untileach sub-dense area contains a single cluster.

As a density-based approach, DHC effectively detects the co-expressed genes (which have rel- atively higher density) from noise (which have relatively lower density), and thus is robust in thenoisyenvironment.Furthermore,DHCisparticularlysuitableforthe“high-connectivity”character- istic of gene expression data, because it first captures the “core” part of the cluster and then divides the borders of clusters on the basis of the “attraction” between the data objects. The two-level hierarchical representation of the data set not only discloses the relationship between the clusters(via density tree), but also organizes the relationship between data objects within the same cluster (via attraction tree). However, to compute the density of data objects, DHC calculates the distance be- tween each pair of data objects in the data set. The computational complexity of this step is , which makes DHC not efficient. Furthermore, two global parameters are used in DHC to control the splitting process of dense areas. Therefore, DHC does not escape from the typical difficulty to determine the appropriate value of parameters.

Sample-based Clustering

Within a gene expression matrix, there are usually several particular macroscopic phenotypes ofsamplesrelatedtosomediseasesordrugeffects,such asdiseasedsamples,normalsamplesordrug treated samples. The goal of sample-based clustering is to find the phenotype structures or sub- structures of the samples. Previous studies [24] have demonstrated that phenotypes of samples can be discriminated through only a small subset of genes whose expression levels strongly correlate with the class distinction. These genes are called informative genes. The remaining genes in the gene expression matrix are irrelevant to the division of samples of interest and thus are regarded as noise in the dataset.

Although the conventional clustering methods, such as K-means, self-organizing maps (SOM),hierarchicalclustering(HC)canbedirectlyappliedtoclustersamplesusingallthegenesasfeatures, the signal-to-noise ratio (i.e., the number of informative genes versus that of irrelevant genes) is usually smaller than , which may seriously degrade the quality and reliability of clustering results[73,63].Thus,particularmethodsshouldbeappliedtoidentifyinformativegenesandreduce gene dimensionality for clustering samples to detect theirphenotypes.

The existing methods of selecting informative genes to cluster samples fall into two major cat- egories: supervised analysis (clustering based on supervised informative gene selection) and unsu- pervised analysis (unsupervised clustering and informative gene selection).

Clustering based on supervised informative geneselection

The supervised approach assumes that phenotype information is attached to the samples, for exam- ple, the samples are labeled as diseased vs. normal. Using this information, a “classifier” which only contains the informative genes can be constructed. Based on this “classifier”, samples can be clustered to match their phenotypes and labels can be predicted for the future coming samplesfrom the expression profiles. Supervised methods are widely used by biologists to pick up informative genes. The major steps to build the classifierinclude:

Training sample selection. In this step, a subset of samples is selected to form the training set. Since the number of samples is limited (less than), the size of the training set is usually at the same order of magnitude with the original size ofsamples.

Informative gene selection. The goal of informative gene selection step is to pick out those genes whose expression patterns can distinguish different phenotypes of samples. For example,ageneisuniformlyhighinonesampleclassanduniformlylowintheother[24].A series of approaches to select informative genes include: the neighborhood analysisapproach [24]; the supervised learning methods such as the support vector machine (SVM) [10], and a variety of ranking based methods [6, 43, 47, 49, 68,70].

Sample clustering and classification. After about [24, 42] informative genes which manifest the phenotype partition within the training samples are selected, the whole setofsamplesareclusteredusingonlytheinformativegenesasfeatures.Sincethefeaturevolume is relatively small, conventional clustering algorithms, such as KmeansorSOM,areusuallyappliedtoclustersamples. Thefuturecomingsamplescanalsobeclassifiedbasedontheinformativegenes,thusthesupervisedmethodscanbeusedtosolvesampleclassification problem.

Unsupervised clustering and informative geneselection

Unsupervised sample-based clustering assumes no phenotype information being assigned to any sample. Since the initial biological identification of sample classes has been slow, typically evolv- ing through years of hypothesis-driven research, automatically discovering samples’phenotypespresentsasignificantcontributioningeneexpressiondataanalysis[24].Asanunsupervisedlearning method, clustering also serves as an exploratory task intended to discover unknown sub-structures in the samplespace.

Unsupervised sample-based clustering is much more complex than supervised manner since no training set of samples can be utilized as a reference to guide informative gene selection. Many mature statistic methods and other supervised methods can not be applied without the phenotypes of samples known in advance. The following two new challenges of unsupervised sample-based clustering make it very hard to detect phenotypes of samples and select informative genes.

Since the number of samples is very limited while the volume of genes is very large, such data sets are very sparse in high-dimensional genes space. No distinct class structures of samples can be properly detected by the conventional techniques (for example, densitybased approaches). Most of the genes collected may not necessarily be of interest. A small percentage (less than [24]) of genes which manifest meaningful sample phenotype structure are buried in large amount of noise. Uncertainty about which genes are relevant makes it difficult to select informative genes.

Two general strategies have been employed to address the problem of unsupervised clustering and information gene selection: unsupervised gene selection and interrelated clustering.

Unsupervised gene selection. The first strategy differentiates gene selection and sample cluster- ing as independent processes. First the gene (feature) dimension is reduced, then the conventional clusteringalgorithmsareapplied.Sincenotrainingsamplesareavailable,geneselectiononlyrelies on statistical models to analyze the variance in the gene expressiondata.

Alter et al. [4] applied the principal component analysis (PCA) to capture the majority ofthevariationswithinthegenesbyasmallsetofprincipalcomponents(PCs),called“eigen-genes.”The samples are then projected on the new lower-dimensional PC space. However, eigen-genes do not necessarily have strong correlation with informative genes. Due to the large number of irrelevant genes, discriminatory information of gene expression data is not guaranteed to be the type of user- interested variations. The effectiveness of applying PCA before clustering is discussed in[75].

Ding et al. [17] used a 1 statistic method to select the genes which show large variance in the expression matrix. Then a min-max cut hierarchical divisive clustering approach is applied toclustersamples.Finally,thesamplesareorderedsuchthatadjacentsamplesaresimilarandsamples far away are different. However, this approach relies on the assumption that informative genes exhibit larger variance than irrelevant genes which is not necessarily true for the gene expression data sets [75]. Therefore, the effectiveness of this approach also depends on the datadistribution.

Interrelatedclustering. Whenwehaveacloserlookattheproblemsofinformativegeneselection and sample clustering, we will find they are closely interrelated.

Once informative genes have been identified, then it is relatively easy to use conventional clustering algorithms to cluster samples.On theotherhand, oncesampleshavebeencorrectlypartitioned,somesupervisedmethodssuchast-test scores and separation scores [68] can be used to rank the genes according to their relevance to the partition. Genes with high relevance to the partition are considered as informative genes. Based on thisobservation, thesecondstrategyhasbeensuggestedtodynamicallyusetherelationshipbetween the genes and samples and iteratively combine a clustering process and a gene selection process. Intuitively, althoughwedonotknowtheexactsamplepartitioninadvance,foreachiterationwecan expect to obtain an approximatepartitionthatisclosetothetargetsamplepartition.The approximate partition allows the selection of a moderately good gene subset, which will, hopefully, draw the approximat epartition evenclosertothetargetpartitioninthenextiteration.Afterseveraliterations,thesamplepartitionwillconvergetothetruesamplestructure,andtheselectedgeneswillbe feasible candidates for the set of informativegenes. Xingetal.[73]presentedasamplebasedclusteringalgorithmnamedCLIFF(CLusteringviaIterativeFeatureFiltering)whichiterativelyusessamplepartitions asareferencetofiltergenes.In [73], non-informative genes were divided into the following three categories: 1)non-discriminative genes (genes in the “off” state); 2) irrelevant genes (genes do not respond to the physiological event);and3) redundantgenes(genesthatareredundantorsecondaryresponsestothebiological orexperimentalconditionsthatdistinguishdifferent samples).CLIFFfirstusesatwocomponentGaussianmodeltorankallgenesintermsoftheirdiscriminabilityandthenselectasetofmostdiscriminantgenes.Itthenappliesagraphtheoreticalclusteringalgorithm,NCut(ApproximateNormalizedCut),togenerateaninitialpartitionforthesamplesandentersaniterationprocess.For each iteration, the input is a reference partition "of the samples and the selected genes. Firstascoringmethod,namedinformationgainranking,isappliedtoselectasetofmost“relevant”genes basedonthesamplepartition".TheMarkovblanketfilteristhenusedtofilter“redundant”genes. Theremaininggenesareusedasthefeaturestogenerateanewpartition"ofthesamplesbyNCut clustering algorithm. The new partition "and the remaining genes will be the input of the next iteration.Theiterationendsifthisnewpartition"isidenticaltotheinputreferencepartition". However,thisapproachissensitivetotheoutliersand noiseofthesamplessincethegenefilteringhighlydependsontheresultoftheNCutalgorithmwhichisnotrobusttothenoiseandoutliers.Tang et al. [64, 65] proposed iterative strategies for interrelated sample clustering and informa- tive gene selection. The problem of sample-based clustering is formulated via an interplay between sample partition detection and irrelevant gene pruning. The interrelated clustering approaches con- tained three phases: an initialization partition phase, an interrelated iteration phase, and a class validation phase. In the first phase, samples and genes are grouped into several exclusive smaller groups by conventional clustering methods K-means or SOM. In the iteration phase, the relation- ship between the groups of the samples and the groups of the genes are measured and analyzed. A representation degree measurement is defined to detect the sample groups with high internal co- herence as well as large difference between each other. Sample groups withhighrepresentationdegreearepostedtoformapartialorapproximatesamplepartitioncalledrepresentative pattern.

The representative pattern is then used to direct the elimination of irrelevant genes. In turn, the remaining meaningful genes were used to guide further representative pattern detection. The termination of the series of iterations is determined by evaluating the quality of the sample partition. This is achieved in the class validation phase by assigning coefficient of variation (CV) to measure the “internally-similar and well-separated” degree of the selected genes and the related sample partition. The formula for the coefficient of variation is: "+ where ! represents thenumber ofsamplegroups, indicates the center sample vector of group , and representsthestandard deviation of group). When a stable and significant sample partition emerges, the iteration stops, and the finial sample partition become the result of the process. This approach delineates the relationships between sample groups and gene groups while conducting aniterativesearchforsamples’phenotypesandinformativegenes.Sincetherepresentativepatternidentifiedin eachstep is only formed by “internally-similar and well-separated” sample groups, this approach is robust to the noise and outliers of thesamples.

Subspace Clustering

The clustering algorithms discussed in the previous sections are examples of “global clustering”; for a given data set to be clustered, the feature space is globally determined and is shared by all

Table 2: Some data sets for sample-based analysis.resultingclusters,andtheresultingclustersare exclusiveandexhaustive.However,itiswellknown in molecular biology that only a small subset of the genes participates in any cellular process of interestandthatanycellularprocesstakesplaceonlyinasubsetofthesamples.Furthermore,asingle gene may participate in multiple pathways that may or may

not be coactive under all conditions, so that a gene can participate in multiple clusters or in none at all. Recently a series of subspace clustering methods have been proposed [22, 11, 40] to capture coherence exhibited by the "blocks" within gene expression matrices. In this context, a "block" is a sub-matrix defined by a subset of genes on a subset ofsamples.

Subspace clustering was first proposed by Agrawal et al. in general data mining domain [1] to find subsets of objects such that the objects appear as a cluster in a subspace formed by a subset of the features. Figure 2 shows an example of the subspace clusters (A and B) embedded in a gene expression matrix. In subspace clustering, the subsets of features for various subspace clusters can be different. Two subspace clusters can share some common objects and features, and someobjects may not belong to any subs pacecluster.

For a gene expression matrix containing genes and samples, the computational complexity of a complete combination of genes and samples is so that the problem of globally optimal block selection is NP-hard. The subspace clustering methods usually define models to describe the target block and then adopt some heuristics to search in the gene-sample space. In the following subsection, we will discuss some representative subspace clustering algorithms proposed for gene expression matrices. In these representative subspace clustering algorithms, genes and samplesare

Figure 2: Illustration of subspace clusters. treated symmetrically such that either genes or samples can be regarded as objects or features.

Coupled two-way clustering(CTWC)

Getz et al. [22] model the block as a stable cluster with features () and objects (), where both andcan be either genes or samples. The cluster is "stable" in the sense that, when only the features in are used to cluster the corresponding , does not split below some threshold. CTWC provides a heuristic to avoid brute-force enumeration of all possiblecombinations.Onlysubsetsofgenesorsample sthatareidentifiedasstableclustersinpreviousiteration sarecandidates for the nextiteration.

CTWC begins with only one pair of gene set and sample set (,), where is the set contain- ing all genes and is the set that contains all samples. A hierarchical clustering method, called the super-paramagneticclusteringalgorithm(SPC)[8],isapplied toeachset,andthestableclustersof genes and samples yielded by this first iteration are and . CTWC dynamically maintainstwo listsofstableclusters(genelist**O**andsamplelist**O**)an dapairlistofpairsofgeneandsample subsets (,). For each iteration, one gene subset from **O** and one sample subset from **O**thathavenotbeenpreviouslycombinedarecoupled andclusteredmutuallyasobjectsandfeatures.Newl ygeneratedstableclustersareaddedto**O**and,andap ointerthatidentifiestheparentpairisrecordedinthe pairlisttoindicatetheoriginoftheclusters.Theiterati oncontinuesuntilnonewclustersarefoundwhichsa tisfysomecriterion,suchasstabilityorcriticalsize. CTWC was applied to a leukemia data set [24] and a colon cancer data set [3]. For the leukemia dataset,CTWCconvergesto49stablegeneclustersand 35stablesampleclustersintwoiterations. Forthecoloncancerdataset,76stablesampleclustersan d97stablegeneclusterswerereportedby CTWC in two iterations. The experiments demonstrated the capability of CTWC to identify sub- structures of gene expression data which cannot be clearly identified when all genes or samples are used as objects orfeatures.

However, CTWC searches for blocks in a deterministic manner and the clustering results are therefore sensitive to initial clustering settings. For example, suppose (,) is a pair of stable clusters. If, during the previous iterations, was separately assigned to several clusters according to features , or was separated in several clusters according to features,then(,)canneverbefoundbyCTWCinthefollo wingiterations.AnotherdrawbackofCTWCisthatclus teringresults are sometimes redundant and hard to interpret. For example, for the colon cancer data, a total of 76 sample clusters and 97 gene clusters were identified. Among these, four different gene clusters partitioned the samples in a normal/cancer classification and were therefore redundant, whilemanyoftheclusterswerenotofinterest,i.e.,hardt ointerpret.Moresatisfactoryresultswouldbeproduced iftheframeworkcanprovideasystematicmechanismto minimizeredundancyandranktheresulting clusters according to significance.

Plaidmodel

The plaid model [40] regards gene expression data as a sum of multiple "layers", whereeachlayermayrepresentthepresenceofaparticul arbiologicalprocesswithonlyasubsetofgenesandasub setofsamplesinvolved.Thegeneralizedplaidmodelisf ormalizedas2 3,wherethe expression level 2 of gene under sample is considered coming from multiple sources. To be specific, **.** is the background expression level for the whole data set, and **.** describes the contribution from layer The parameter (or 3) equals when gene (or sample) belongstolayer , and equals otherwise. Theclusteringprocesssearchesthelayersinthedatas etoneafteranother,usingtheEMalgorithmtoestimat ethemodelparameters.Supposethefirst **!**

layershavebeenextracted,the!)4 layer is identified by minimizing the sum of squared errors 56.3 , where 623is the residual from the first ! layers. Theclustering process stops when the variance of expression levels within the current layer is smaller than a threshold.

The plaid model was applied to a yeast gene expression data set combined from several time- series under different cellular processes [40]. Totally, 34 layers were extracted from the data set, among which interesting clusters were found. For example, the second layer was recognized as dominated by genes that produce ribosomal proteins involved in protein synthesis in which mRNA is translated. However, the plaid model is based on the questionable assumption that, if a geneparticipatesinseveralcellularprocesses,thenitsexpressionlevelisthesumofthetermsinvolvedin the individual processes. Thus, the effectiveness and interpret ability of the discovered layers need further investigation.

REFERENCES

[1]. Agrawal,R.,Gehrke,J.,Gunopulos,D.,andRaghavan,P.Automaticsubspaceclusteringofhigh dimen- sional data for data mining applications. In SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, pages 94–105,1998.

[2]. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M. et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature, Vol.403:503–511, February2000.

[3]. Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonu- cleotide array. Proc. Natl. Acad. Sci. USA, Vol. 96(12):6745–6750, June1999.

[4]. Alter O., Brown P.O. and Bostein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA, Vol. 97(18):10101–10106, Auguest2000.

[5]. Ankerst,Mihael,Breunig,MarkusM.,Kriegel,Hans-Peter,Sander,Jrg.OPTICS:OrderingPointsTo Identify the Clustering Structure. Sigmod, pages 49–60,1999.

[6]. Ben-DorA.,FriedmanN.andYakhiniZ.Classdiscoveryingeneexpressiondata.InProc.FifthAnnualInter.Conf.onComputationalMolecularBiology(RECOMB2001),pages31–38.ACMPress,2001.

[7]. Ben-Dor A., Shamir R. and Yakhini Z. Clustering gene expression patterns. Journal of Computational Biology, 6(3/4):281–297,1999.

[8]. Blat, M., S. Wiseman and E. Domany. Super-paramagnetic clustering of data. Phys. Rev. Letters, 76:3251–3255,1996.

[9]. Brazma, Alvis and Vilo, Jaak. Minireview: Gene expression data analysis. Federation of European Biochemical societies, 480:17–24, June2000.