DNA Microarrays and Computational Analysis of DNA Microarray Data in Cancer Research

Mario Medvedovic, Jonathan Wiest

Abstract

The advent of DNA microarray technology has added a new dimension to the field of molecular carcinogenesis research. DNA microarrays have been used as a tool for identifying changes in gene expression and genomic alterations that are attributable to various stages of tumor development. Patterns defined by expression levels of multiple genes across different types of cancerous and normal tissue samples have been used to examine relationships between different genes, and as the tool for molecular classification of different types of tumor. The analysis of relatively large datasets generated in a typical microarray experiment generally requires at least some level of computer-aided automation. On the other hand, the large number of hypotheses that are

implicitly tested during the data analysis, especially when identifying patterns of expression through supervised and unsupervised learning approaches, require careful assessment of statistical significance of obtained results. These basic requirements have brought to the fore front the need for developing statistical models and corresponding computational tools that are specifically tailored for the analysis of microarray data. Such models need to be able to differentiate between faint, yet statistically significant and biologically important signals, and patterns that are generated by random fluctuations in the data. In this endeavor, it is important to keep in mind the abundance of already existing statistical and machine learning methodologies which can serve as the starting point for developing more specialized techniques. Here we describe different uses of DNA microarray technology in molecular carcinogenesis research and related methodological approaches for analyzing and interpreting DNA microarray data obtained in such experiments.

1) Introduction

Novel molecular biology technologies for performing large numbers of biological measurements in parallel provide an unprecedented opportunity for uncovering the molecular basis of cancer and mechanisms of cancer induction by carcinogens. The large volume of data generated by experiments utilizing such assays, as well as relatively high experimental noise often associated with them require a careful statistical/computational analysis. The most prominent of such novel technologies are DNA microarrays, which facilitate the assessment of a whole transcriptome of a cell population in single experiments.

DNA microarrays are glass slides on which a large number of DNA probes, each corresponding to a specific mRNA species, or a genomic DNA region, are placed at predefined positions. DNA probes are either synthesized in-situ [1,2], or they are pre-synthesized and then spotted on the slide [3]. Two most commonly used technologies are Affymetrix in-situ synthesized microarrays and the spotted microarray technology developed at Stanford University. In gene expression experiments RNA is extracted from the biologic sample, reverse transcribed into cDNA and fluorescently labeled. Such labeled cDNA representing the transcriptome of the biologic sample is then hybridized on the microarray. The amount of the labeled cDNA that hybridizes to each probe on the microarray is proportional to the relative abundance of the corresponding cDNA. The expression of all genes is then quantified by measuring the intensity of the dye used to label the RNA. The most common experimental protocol used with spotted microarrays consists of labeling two RNA extracts with different dyes and co-hybridizing the samples to the same microarray (two-channel microarrays). While this approach introduces some restrictions on the experimental design [4], the overall principles of the two major technologies are the same.

Quantification of individual gene expressions proceeds by various normalization procedures whose role is to remove systematic biases and to rescale measurements on different arrays to be directly comparable. The development of an appropriate normalization procedure is still an active research topic [5-7]. Normalized data is used to identify genes differentially expressed in different tissues, to identify groups of genes with similar pattern of expression across different biological states and to construct rules for classifying different biological samples based on their expression profiles.

In general, computational analysis of microarray data can be separated into the single-gene at a time analysis, in which the data for each gene is analyzed independently of the data for any other gene, and multiple-gene at a time analyses in which the data for all or a sub-group of genes is jointly analyzed. In a single gene at a time analysis, the goal is generally to identify genes that are differentially expressed in different tissues. In a multiple gene at a time analysis, the information from multiple genes is combined to identify global patterns of expression that can offer additional insights not available by looking at genes separately.

2. Applications of microarrays

Although the most common application of microarrays is in monitoring gene expression, the other extremely relevant application in the context of cancer research is the microarray based Loss of Heterozygosity (LOH) analysis [8,9].

Gene expression data generated using microarrays is generally used to identify genes that are differentially expressed under different experimental conditions, identify groups of genes with similar expression profiles across different experimental conditions (co-expressed genes) and classify the biologic sample based on the pattern of expression of all or a subset of genes on the microarray. Differentially expressed genes as well as groups of co-expressed genes can be used to hypothesize which pathways were involved in a particular biologic process. Additionally, clusters of co-expressed genes can be used to hypothesize the functional relationship of a clustered gene, and as a starting point of dissecting regulatory mechanisms underlying the co-expression. In the context of tumor classification, gene expression profiles have been used as complex biomarkers defining the tumor as well as different sub-classes of tumor.

Genomic instability is central to the development of cancer. Gene amplifications and deletions are a major factor in tumorigenesis. Copy number changes are important in the understanding of cancer biology, diagnosis, and progression. These genetic alterations can lead to expression changes in oncogenes or tumor suppressor genes, respectively. Changes in gene expression as a result of these alterations are likely to be the driving force behind many of the amplifications and deletions that occur giving the transformed cell a growth advantage.

Comparative genomic hybridization (CGH) is a technique that analyzes the global genetic alterations in cells. The procedure detects both deletions and amplifications of the genome and allows for the global analysis of genetic alterations in tumors. The traditional CGH uses differentially labeled test and reference genomes that are co-hybridized to normal metaphase chromosome spreads. The fluorescent ratio of the labeled DNAs is then measured over the length of the chromosomes to determine regions of gain or loss. Analyzing the data indicates amplified and deleted regions of the genome based on the intensity of each fluorescent signal. Analyzing numerous samples demonstrates the frequency of the genomic aberrations. One disadvantage of the technique is that the sizes of the alterations need to be fairly large, for example on the order of 5 to 10 megabases, to be detected [10]. An additional problem is the procedure is very labor intensive and not amenable to analyzing large numbers of samples. Other methods, including microsatellite marker analysis and fluorescent *in situ* hybridization provide a higher resolution map, but are also labor intensive and may not be applicable to whole genome analysis.

The advantages of newly developed microarray based CGH assays are numerous. This technology is theoretically capable of assessing relatively small genomic aberrations and is capable of the high-throughput analysis. A clear demonstration of the improved resolution of the microarray CGH over the traditional approach was offered in microarray-based CGH analysis of the SKBR3 breast cancer cell line [11]. In these experiments, microarray-based CGH analysis improved the resolution of amplicons in the 8q regions over the traditional analysis [12].

3. Analysis of gene expression microarrays

At this point we assume that the quantification of flourescence intensities of individual spots on the microarray has been completed. The analysis of microarray data in most situations proceeds by normalizing data, identifying genes whose expression changes between different different experimental conditions and performing multivariate analyses, such as clustering and classifying.

a. Data normalization

The first step in the computational analysis of microarray data almost always consists of performing various transformations with the aim of reducing systematic variability. Although the optimal procedures are still being developed, a certain concensus is emerging on the appropriate ways to perform initial data normalizations in microarray experiments [13]. In the case of the spotted two-channel arrays, two major sources of the systematic variability are the spot-specific local background fluorescence and the difference in the overall intensities of the two fluorescent dyes (Cy 3 and Cy5). The process of normalization generally proceeds by subtracting the local background and centering the log-ratios of two channel intensities around zero. In Figure 1, log-ratios of

background-subtracted intensities in two channels are plotted against their average. The line describing the average behavior of data is the local regression (loess) curve [14].

(Insert Figure 1 here)

Initially, the common practice was to center the log-ratios by subtracting the overall median value. However, it is fairly obvious from the Figure 1 that such an adjustment is likely to "over-adjust" high-intensity spots and "under-adjust" the low intensity spots. It turns out that the local regression based normalization which subtracts the fitted loess curve value from the corresponding log-ratio generally does a better job of reducing this channel bias [15,16] and is gaining wide acceptance. In the case of the Affymetrix data, similar strategy of scaling data on all microarrays in the experiment to a "control" chip so that all chips have equal median intensities has been commonly used. Recently, alternative approaches based on the intensity-specific normalizations have been introduced as well [17].

b. Detecting differentially expressed genes across different samples

The purpose of the statistical analysis in the process of identifying differentially expressed genes is to assess the reproducibility of observed changes in gene expression by assessing their statistical significance. This is done by comparing the magnitude of the observed changes in gene expression to the magnitude of random fluctuations in the data. For example, in the traditional t-test analysis, the average differential expression observed in replicated experiments is divided by its standard error and the obtained quantity (t-statistic) is compared to its theoretical distribution under the assumption that the observed average differential expression is a result of random fluctuations in the data. In the context of the cancer-related microarray data, paired t-test and the step-down

Bonferonni adjustment was used to identify genes whose expression is affected in testis of mice that were gestationaly and lactationaly exposed to diethylstilbestrol [18]. Furthermore, identifying genes that are differentially expressed between different classes of tumor tissues is often a first step in identifying relevant genes for the purpose of cluster analysis and tumor classification [19,20].

For any kind of analysis to be successful in assessing reproducibility of observed results, it is necessary to apply an appropriate experimental design in the process of gathering data. The key requirements for the appropriate experimental design are that it addresses all relevant sources of variability. Suppose that we want to identify genes that are differentially expressed between two different types of tumors using two-channel spotted microarrays. The logical requirement for the implicated genes is that they are on average differentially expressed between the two tissue types. Several decisions that are made prior to performing experiments are going to significantly impact the reproducibility of the results of the experiment regardless of the subsequent statistical analysis. Assuming that we performed appropriate normalization of the data, two unavoidable types of variability will be present in the data. One is the technical variability that is introduced in the process of isolating and labeling RNA, fabrication of microarrays, scanning process, etc. The other is the biological variability between the different tissue samples of the same kind used in the analysis.

In most biological applications, the biologic variability dominates the technical variability. For example, the variability between different measurements of the same tissue sample will be much smaller than the variability between different tissue samples of the same kind (e.g. same tumor type from multiple individuals). To reduce the overall

variability in our hypothetical experiment, one could be tempted to use only one tumor sample of each kind and perform several technical replicates. Due to the lower variability than if different tumors are used in replicated experiments, such an approach is likely to result in more genes being pronounced differentially expressed. The obvious problem is though, that such results will not generalize to the whole population of these two types of tumors and consequently will not be reproducible.

On the other hand, there are several sources of variability that can be efficiently removed from the estimates of differential expression using the factorial Analysis of Variance (ANOVA) approach. Two such sources that are more commonly addressed in the statistical analysis of microarray data are gene-specific dye effects and the array effect. These effects are manifested in the fact that fluorescence measurement of one dye are reproducibly higher than the other dye in gene-specific fashion meaning that the effect varies from one gene to another. If this source of variability is not taken into account when the experiment is being designed, it could result in falsely implicating non-differentially expressed genes as well as in missing truly differentially expressed genes. One way to deal with this problem is to perform “dye-flips”, meaning that different RNA samples from the same tissue type are labeled with different dyes. If the number of replicates labeled with Cy3 is equal to the number of replicates labeled with Cy5, this will remove the systematic bias from the analysis. However, if this new source of variability is not extracted in the ANOVA analysis, it can seriously inflate the variability of the differential expression estimates.

In the factorial ANOVA one estimates contributions of different systematic sources of variability and extracts them from the estimates of the effect of interest. For example

the simplest linear model that allows for the extraction of the gene-specific dye effect using ANOVA is

$$Y_{ijk}=\mu + T_i + D_j + A_k + \varepsilon_{ijk}$$

Where Y_{ijk} is the expression measurement on the k^{th} microarray of the tissue type i labeled by the j^{th} dye (j=1 for Cy3 and j=2 for Cy5). μ is the overall expression level for this gene, ε_{ijk} is the random error in Y_{ijk} unexplained by factors in the model, and T_i is the effect of the i^{th} tissue type on the expression level adjusted for the dye effect (D_j) and microarray (A_k) measuring the differential expression of the gene between different tissue types. By estimating differential expression after adjusting for the dye effect, one effectively removes the variability introduced by "flipping dyes" from the analysis. A more thorough review of experimental design issues in microarray experiments can be found elsewhere [21]. Issues relating to using ANOVA in analyzing microarray data are discussed in the context of the fixed-effect model [22], and in the context of the mixed-effect model [23].

Statistical methods for identifying differentially expressed genes have come a long way from initial heuristic attempts [24], through the realizations that rigorous statistical analysis of replicated data is needed [25], to sophisticated statistical modeling using frequentist and Bayesian approaches [26]. Generic statistical methods of the analysis of variance [22] and mixed models [23] are complemented with specialized maximum likelihood approaches [27], and Bayesian analysis flavored approaches [28-30]. While the consensus about the optimal method has still not been reached, the intense statistical research is a promising sign.

One of the most daunting issues in the process of identifying differentially expressed genes is the severe problem of multiple comparisons. Presently, expression level of up to more than 20,000 different genes can be assessed on a single microarray. Searching for genes whose expression change is statistically significant corresponds to testing 20,000 hypotheses simultaneously. If each of these tests is performed at the commonly used significance level of alpha=.05, meaning that we expect for 5% of genes that are not differentially expressed to be falsely implicated, we expect on average 1,000 falsely implicated genes. The simplest way to deal with this multiple comparison problem is to divide the significance level by the number of hypotheses testing (Bonferonni adjustment). In the case of 20,000 hypotheses, this will mean that individual hypotheses will be tested at the significance level of alpha=.0000025. Such a level is virtually unattainable in simple experiments with few experimental replicates. While the multiple comparison issue cannot be avoided, a better balance can be struck between the need to avoid false positives and false negatives. The False Discovery Rate (FDR) adjustment [31] keeps the balance between the specificity and the sensitivity of microarray data analysis [30]. In contrast with traditional adjustments that control the probability of a single false positive in the whole experiment, the FDR approach controls the proportion of false positives among the implicated genes. For example, if 20 genes are selected using FDR=.05, one of them will on average be a false positive regardless of the total number of genes. The traditional (e.g., Bonferonni) adjustment will limit the probability of a single false positive to .05, resulting in a possibly conservative testing procedure.

c) Identifying clusters of co-expressed genes

i) Overview of clustering approaches

The high-dimensional nature of microarray data has prompted the widespread use of various multivariate analytical approaches aimed at identifying and modeling patterns of expression behavior. In cancer research, cluster analysis has been commonly utilized to identify groups of genes with a common pattern of expression across different tissues as well as to group tissues with similar genetic expression profiles. Results of the cluster analysis have been used to infer common biologic function and the co-regulation of co-expressed genes in response to mutagenic treatments [32] and P53-specific DNA damage response [33], to identify genes groups of genes whose pattern of expression can serve as the marker of various stages in tumor progression [19] and to assess the possibility of classifying different kind of tumors based on their gene expression profiles [34-37]. Relevance of different clusters obtained by hierarchically clustering breast carcinomas was confirmed by correlating them with the mutational status of the P53 gene and the clinical outcome [20].

(Insert Figure 2 here)

The power of the clustering approach in interpreting patterns of expression of groups of genes is demonstrated in Figure 2. The data in Figure 2 comes from the publicly available yeast cell-cycle dataset [38]. If we just observe expression patterns of two genes in Figure 2A, about whom we know very little, we would conclude that their expression profiles are highly correlated. This might lead us to conjecture that these two genes are participating in the same crucial point of the cell cycle progression. On the other hand, if we knew that there are two different expression patterns that our to genes could be

associated with, given in Figure 2B, we would probably conclude that these two genes are actually representative of two different expression patterns. If we don't know of the existence of such two patterns but are given 74 genes that define these to patterns, a simple hierarchical clustering procedure will easily identify two clusters in Figure 2C and 2D, defining two patterns in Figure 2B and associating our two genes to distinct clusters. Since the advent of the microarray technology virtually all traditional clustering approaches have been applied in this context and numerous new clustering approaches have been developed.

(Insert Figure 3 here)

Hierarchical clustering procedures were the first to be applied in the analysis of microarray data [39] and are still the most commonly used clustering procedure in this context. Such methods rely on the calculation of pairwise distances or similarities between the gene profiles. Various correlation coefficients are the most commonly used measures of similarity. Hierarchical agglomerative methods generally proceed by grouping genes and groups of genes based on such pairwise measures of similarity. In this process, the distance between two groups of genes is calculated as a function of individual pairwise distances of genes in two groups using different "linkage" functions. "Single-linkage" corresponds to the minimum pairwise distance between genes in two different groups, "complete-linkage" corresponds to the maximum distance, and "average-linkage" corresponds to the average distance [40]. Virtually every publication related to utilizing microarrays for gene expression profiling of tumor tissues and cell lines contains a figure with genes and/or tissues organized in this fashion.

Partitioning approaches, on the other hand, work by iteratively re-assigning profiles in a pre-specified number of clusters with the goal of optimizing an overall measure of fit. Two of the most commonly used traditional approaches are k-means algorithm and self-organizing map (SOM) method, first applied in this context by Tavazoie *et al.* [41] and Tamayo *et al.* [42], respectively. One of the problems with clustering methods that are based on pairwise distances of expression profiles is that, at least in the initial steps, only data from two profiles is used at a time. That is, the information about relationships between the two profiles and the rest of the profiles is not taken into account although these relationships can be very informative about the association between the profiles. The major drawback of partitioning approaches is the need to specify the number of clusters. For example, given that we know that there are 2 clusters in the data in Figure 3, both k-means and SOM's will uncover the two clusters of interest. However, while in the hierarchical structure in Figure 3, it is immediately obvious that there are two clusters of data, both k-means and SOM's require this to be known prior to the analysis.

In a model-based approach to clustering, the probability distribution of observed data is approximated by a statistical model. Parameters in such a model define clusters of similar observations and the cluster analysis is performed by estimating these parameters from the data. In a Gaussian mixture model approach [43], similar individual profiles are assumed to have been generated by the common underlying "pattern" represented by a multivariate Gaussian random variable [44-45]. In the situation where the number of clusters is not known, this approach relies on ones ability to identify the correct number of mixture components. A mixture based method for clustering expression profiles that produces clusters by integrating over models with all possible number of clusters was

developed [46]. In this approach, the joint distribution of the data is modeled by a specific hierarchical Bayesian model and the posterior distribution of clusterings is generated using a Gibbs sampler.

Model-based clustering procedures have been shown to have desirable properties in various comparative studies examining properties of different clustering procedures [46,47]. Recently, finite mixture models as implemented in the AutoClass software package [48] were used to refine the clustering of gene expression profiles of human lung carcinomas produced by the hierarchical procedures [49]. A similar method was applied to identify genes related to malignancy of colorectal carcinomas [50].

ii) Assessing statistical significance of observed patterns

A reliable assessment of reproducibility of observed expression patterns and gene clusters is one of the burning issues in cluster analysis. Since cluster analysis has generally been used as an exploratory analysis tool, establishing statistical significance of observed results has not been a priority. However, just like in the case of establishing the statistical significance of differential expressions, an assessment of the reproducibility of observed patterns is necessary before one can take them seriously. Unfortunately, establishing the statistical significance of different features of observed clusters is a much more difficult problem than establishing differential expression of individual genes. Two exceptions are the significance of the existence of the overall clustering structure and the significance of pairwise associations between individual profiles. However, even the pairwise association between individual profiles is a difficult problem if one assumes possible but unknown clustering structures.

(Insert Figure 4 here)

This can be illustrated in the analysis of the two genes of interest in Figure 2. Suppose we are asking the question whether or not these two genes are co-expressed. Or, in other words, do expression profiles of these two genes belong to the same underlying pattern of expression? If we use correlation as a measure of similarity of these two profiles, the Pearson's correlation turns out to be equal to 0.83. In the context of randomly chosen pairwise correlations for all genes in this dataset this turns out to be statistically significant. As a result, we could be tempted to say that these two genes are co-expressed. However, if we analyze the whole group of genes using hierarchical clustering (Figure 2 and 3), it seems that the two genes belong to two distinct patterns of expression. In this respect some of the newly developed statistical approaches offer a glimmer of hope. For example, in the Bayesian Infinite Mixture Model approach, the posterior probability of any particular clustering feature (say gene1 and gene2 are co-expressed) can be directly assessed from the output of the Gibbs sampler [46]. Such a model-based approach is capable of producing an objective measure of confidence in any such feature after incorporating sources of uncertainty in the process of clustering microarray data (i.e. experimental variability and unknown number of clusters).

When we apply the Bayesian Infinite Mixture (BIM) in the context of two genes in Figure 1, the result is rather unambiguous. First of all the posterior distribution of "distances", which are in this context defined as 1-Posterior Probability of Co-expression, indicates strongly that there are actually two clusters in the data (Figure 4). Furthermore, the posterior probability of the feature of interest, which is that these two genes are co-expressed, after averaging over models with all possible number of clusters, is equal to 0 indicating that data actually offers strong evidence that these two genes are not co-

expressed. The higher precision of the model based on posterior probabilities calculated from the BIM is illustrated by comparing distributions of between- and within-cluster distances for the two clusters in Figure 3 obtained by simple correlation and base on BIM model.

d) Gene expression based tumor classification

Classification of tumor samples based on gene transcription profiling has been one of the earliest and one of the most promising areas of microarray technology applications in cancer research. The concept of using the gene expression profiles as complex markers in classifying different types of caners has been initially demonstrated by classifying different types of acute leukemias [51] and distinguishing between the tumor and normal colon tissues [36]. This approach has also been shown to have a great potential for clinical applications in the areas of tumor classification and toxicity screens of potential drug compounds [52-54].

In general, a "classifier" is a mathematical formula that uses as input values of distinct features of an object and produces an output that can be used to predict to which of the predefined classes the object belongs. In terms of the gene expression data based tumor classification, the objects are tissue samples and the features are genes and their expression levels. The construction of a classifier generally proceeds by selecting an informative set of genes that can distinguish various classes, choosing an appropriate mathematical model for the classifier and estimating parameters of the model based on the "training set" of tissues whose classification we know in advance. Finally, the specificity and the sensitivity of the classifier is tested on the data that was not used in the process of constructing the classifier.

The simplest classifier one can envision consists of a single gene and a cut-off value x_c such that a sample is classified in one class if the expression level of this gene is smaller than x_c and in the other class if it exceeds x_c. In the case when multiple features/genes are used, measurements from all of them are again summarized into a single number by using a variety of multivariate models. Such a summary value is then used in a similar fashion as one would use a single gene value. A hypothetical example of advantages of using expression levels of multiple genes for classifying "metastatic" and "non-metastatic" tumors is depicted in Figure 5. While expression of any single of the two hypothetical genes in Figure 5 are not sufficient for reliably predicting whether the sample is "metastatic" or "non-metastatic" (Figure 5C and 5D), their combination constructed by subtracting the expression of the Gene2 from the expression of Gene1 can separate the two classes almost perfectly (Figure 5E).

Various approaches to selecting informative genes can be coarsely grouped in methods that assess the classification abilities of a single gene at a time and methods that choose groups of genes based on their joint ability to distinguish between different tumor classes. The most common methods in practice to date have been based on choosing genes in one-gene at a time fashion based on the statistical significance or the magnitude of their differential expression between different classes of tumors [19,20]. The combinatorial explosion of possible number of different groups of genes generally makes the second approach of choosing groups of genes based on their joint discriminative capacity very difficult. Comparing all possible sub-groups among 20,000 different genes is clearly impossible. Alternatives to the exhaustive comparison are heuristic optimization techniques such as Genetic Algorithm [55] or constructing groups of gene in a

step-wise fashion. Both of these approaches will not necessarily identify the optimal group of genes but have been shown to often perform quite well in this context.

Mathematical models that have been used so far in constructing tumor classifiers can generally be divided in non-parametric methods such as k-nearest neighbor (KNN) [55], Fisher's linear discriminant analysis (FLDA) [56], and support vector machines (SVM) [57] and the methods based on the statistical model for data in individual classes such as various Gaussian model based classifiers, logistic regression [58], and artificial neural networks (ANN) [59]. Excellent descriptions and introductions into various classification approaches are given elsewhere [60,61] .

(Insert Figure 5 here)

In a KNN classifier for the two-classes situation, the distance of the sample expression profile of the sample to be classified from individual profiles of all training data is first calculated. The sample is then classified to the class having the most members within the k-closest neighbors of the sample. Fisher's linear discriminant function is based on identifying the direction in the k-dimensional space (where k is the number of genes used for the classification) that separates the two classes the best in the sense that it maximizes the ratio of between-classes and within-classes variability. SVM classifiers are based on the idea of the "optimal separating hyper-plane" in the k-dimensional space. It chooses the hyper-plane so the distance from the hyper-plane to the closest point in each class is maximized. For example, the linear (hyper-plane) SVM when k=2 (two-genes classifier) will select the straight line such that the traditional Euclidian distance between the line and the closest points in two classes is maximized. In this sense the linear SVM classifiers are similar to the Fisher's linear discriminant

function based classifier except that the two methods use different criteria to select the separating hyper-plane. ANN-based classifiers will generally fit a non-linear hyper-plane to separate two classes of objects (tissues in our case). Probabilistic models based classifiers generally proceed by estimating the distribution of the features of the classes of objects to be classified. The classification is then based on the identification of the most likely of such distribution that have had generated the sample to be classified. In the hypothetical example in Figure 5, all above mentioned procedures will likely perform very well. Strictly speaking the linear discriminator depicted in Figure 5A, 5B, and 5C corresponds to the Fisher's linear discriminant.

Validation of the predictive accuracy of any particular classifier is an essential step in the classification analysis. The optimal way of validating a classifier's performance is to test it on the samples that were not used in any way in the process of building the classifier. A commonly used strategy is to perform a "leave-one-out" analysis in which each of the samples is left out in the process of building the classifier and then used to test its predictive ability. Average predictive ability of the classification procedure can then be summarized as the proportion of the correctly classified samples in the "leave-one-out" analysis. Ideally, predictive ability is then compared to the predictive ability of the equivalent classifier on the randomized data. This is particularly important when we have different number of samples in different classes. For example, a trivial classification rule of always classifying samples into a single class will have 90% correct predictive rate if 90% of the samples that are being classified come from this class.

How to identify the best set of genes as well as questions related to the optimal mathematical model for constructing classifiers are two of the intense research areas of

computational biology. Results of a comparison study of several traditional classification methods in relation to microarray data based tumor classification [62] suggest the need to base classifiers on statistical theory. For example, it was shown that the maximum likelihood-based classifier clearly outperforms a popular heuristic equivalent [63].

An alternative approach to generating optimal classification features is to use some of the dimension-reduction techniques. The most commonly used method is the Principal Component Analysis (PCA). In the PCA analysis, one seeks a small number of linear combinations of the initial features that in a sense condense the predictive information of the whole set of features. Linear combination of k values $(x_1,\ldots,x_k)$ is defined as $a_1x_1+a_2x_2+\ldots+a_kx_k$ where $(a_1,\ldots,a_k)$ are corresponding linear coefficients. PCA identifies linear combinations of features that maximize the variability between different objects. While this heuristic argument behind PCA works in many situations, it some situations it fails completely. For example, in our hypothetical example in Figure 5, the linear combination with the maximum variability is approximately equal to Gene1+Gene2 and it actually results in worse separation than any of the original variables (Figure 6). Another related dimension-reduction technique is the Partial Least Squares (PLS) method [64] which extends the PCA approach to incorporate the information about the correct classification in the process of identifying optimal linear combinations. Because the method chooses linear combinations that accentuate the relationship between the features and the classification of the training object, it generally results in a better classifier. Generally, use of such procedures in the tumor classification setting has an intuitive appeal that one can use information from a large number of genes without experiencing

problems with the classification methods that perform best when the number of samples is significantly larger than the number of features used by the classifier

4. Analysis of CGH microarray data

The computational analysis of microarray CGH data is to a large extent similar to the analysis of gene expression arrays. The data still needs to be normalized, and the changes in copy numbers of different DNA regions represented on the microarray needs to be established by performing the statistical analysis. Similarly, as in the case of expression arrays, data can be clustered to identify patterns of common amplifications and deletions across all tissue samples. CGH microarray data can be also used to design classifiers in exactly the same way as described for the gene expression data.

One specific feature that distinguishes this type of data is the correlation introduced by the linear organization of genome that can be utilized to improve the sensitivity of such analyses. The basic premise of such analysis is that the closer two DNA regions are genomically, the more likely it is to that if one of them is affected by a gross genomic aberration, the other one will be affected as well. One way to explore such correlations is to use moving average estimates of fluorescence intensities of different DNA probes. Moving averages are calculated by averaging intensities of DNA probes corresponding to several neighboring DNA loci. The amount of “smoothing” induced by such a strategy is dependent on how many neighboring loci are averaged. Since such averages have a potential of completely concealing genomic aberrations covered by a single probe, one has be careful about using it. Presumably, such an analysis can be used within a battery of different analytical approaches with performing experimental replicates still being the preferred approach to reducing variability in fluorescence measurements.

5. Integrating current knowledge and various types of experimental data

Integration of the current knowledge with the new experimental data is done every time a biologist interprets results of a new experiment. Interpreting results of a microarray experiment that can yield hundreds of thousands data points in the traditional informal way can be rather difficult. In this situation, one is forced to limit her/his attention to a subset of genes that were indicated in the initial statistical analysis. However, the sensitivity of the statistical analysis can be critically affected by the incorporation of the prior knowledge. For example, if one can make assumptions concerning the subset of genes most likely to be affected, this subset can be analyzed separately with higher statistical power due to fewer hypotheses that are being tested. Formal methods of integrating accumulated knowledge and information in the analysis are being developed. An example of such methods is the method for scoring likelihood of whole pathway involvement in the process under investigation based on integrating the analysis of expression levels of genes involved in the pathway and the existing pathway information [65]. Similarly, benefits of integrating genomic, functional genomic and proteomic data have been demonstrated [66-69]. For example, a weak evidence of co-regulation implied by co-expression identified in a cluster analysis can be strengthened by the result from a two-hybrid assay that indicated the two corresponding proteins interact or by the shared regulatory elements in their promoter region. Statistical models capable of integrating such diverse data types have been proposed by several investigators [70-73], while the use of joint proteomic and functional genomic data after perturbing a biologic system to reverse engineer the underlying network of molecular

interactions, in the context of the “systems biology” paradigm, has been demonstrated [74,75].

a. From co-expression to co-regulation

Transcriptional regulation is one of the crucial mechanisms used by a living system to regulate protein levels. It is estimated that 5-10% of the genes in eukaryotic genomes encode transcription factors that are dedicated to the complex task of deciding where, when and which gene is to be expressed. Mechanisms applied by these factors range from the recruitment and the activation of the transcriptional pre-initiation complex to necessary modulations of local chromatin structure. Two major determinants of gene expression specificity seem to be the composition of their *cis*-regulatory modules, and the presence/absence and phosphorylation status of *trans*-acting regulatory factors that interact with DNA regulatory modules and each other. However, the exact nature of the interactions between various components of the regulatory machinery is still largely unknown. Identification of co-expressed genes by a cluster analysis of gene expression profiles has often been utilized as a first step in identifying factors regulating expression of different genes. On the other hand, using information about presence of known regulatory elements can be applied to refine the cluster analysis of expression profiles and the simultaneous identification of known regulatory elements causing such co-regulation .

An indirect indication of co-regulation of co-expressed genes is the tendency of co-expressed genes to participate in the same biologic pathway as well as their tendency to code for proteins that interact with each other. In both of these situations, the mechanism of co-regulation might not be at the level of common *cis*-regulatory elements, yet the

need for co-regulation and the actual presence of them is obvious. Actually, it has been shown that the particular expression regulatory mechanism can sometimes be a better determinant of the protein function than even its three dimensional structure. All these suggest that analytical methods capable of integrating information about regulatory sequences, biologic pathways and protein-protein interactions, and expression data generated in microarrays experiments will be better able to create biologically meaningful clusters of genes than clustering expression data alone.

b. Integrating microarray CGH and expression data

In terms of cancer research, tumors represent a naturally perturbed genomic system. Concurrent genomic and functional genomic investigations of tumors by the high-throughput microarray approaches can be used to dissect genetic networks involved in the process of tumorigenesis. In this context, microarrays can be used to both characterize the genomic aberrations and the gene expression in different tumors. Several models mentioned before are capable of integrating such information into a single powerful analysis.

The next logical step in high-through-put analysis is to combine the cDNA array analysis with CGH analysis of tumor samples. This has been accomplished in several recent reports. Fritz et al. [76] applied CGH and cDNA based arrays to liposarcomas and found that tumor subtypes revealed more effectively by clustering genomic profiles than by clustering expression profiles. Weiss et al. [77] have shown that in gastric adenocarcinomas, microarray analysis of genomic copy number changes can predict the lymph node status and survival outcome in patient samples. In breast tumors, Pollack et al. [11] found that 62% of highly amplified genomic regions contain over expressed genes

and in general that a 2-fold change in copy number corresponds to a 1.5-fold change in mRNA levels as detected on the cDNA arrays. Additionally, they report that 12% of all gene expression changes in breast tumors are directly attributed to changes in gene copy number.

In all these reports, the integration of CGH and gene expression data has been achieved by analysing them separately and correlating results of individual analysis. However, it is likely that unified analysis strategies, akin to already mentioned statistical models for joint analysis of expression and regulatory sequence data, will prove beneficial in this context as well.

c. Modeling genetic networks

The ultimate goal of integrating different types of experimental data and current biological knowledge into a mathematical framework is the construction of genetic network models that will help us understand and predict the global dynamics of complex biological processes that define a living cell. The traditional molecular biology approach to characterizing roles of different cellular components has been to collect information on the single gene, single protein or single interaction at a time. However, some characteristics of behavior of the complex network of biochemical interactions defining the living system are unlikely to be recovered by such local approaches [74,78]. For example, the functional role of a gene whose expression is regulated by several transcription factors cannot be fully understood without simultaneously monitoring for the presence and/or activation status of all of them. The ability of DNA microarrays to generate at the same time measurements on a large number of molecules participating in such a network allows for assessing interactions of a substantial portion of the global

network. The complete strategy for such analysis consists of a mathematical model describing interactions of various components of the network, experimental approaches to perturb the network, biologic assays for quantitating the effects of such perturbation and the inference procedures for estimating parameters of the assumed model [79].

Mathematical models describing dynamics of biochemical networks include the deterministic ordinary differential equation (o.d.e.) based models of kinetics of coupled chemical reactions [80,81], stochastic generalization of such models following the Gillespie's algorithm [82] for simulation approach to the chemical master equations [83-86], Boolean network models which reduce the information about the abundance of various interacting molecules to a binary variable representing on/off (0/1, present/absent) states [87], and Bayesian networks [88] and probabilistic graphical models in general [89]. All of these mathematical models have certain advantages and disadvantages depending on the goal of the analysis, available data and the knowledge about interactions of various molecules in the network. A thorough overview of these models in the context of genetic regulatory networks can be found elsewhere [90].

The specification of an o.d.e. model requires detailed knowledge of the modeled interactions and is intended for examination of the overall dynamics of the system when individual relationships between components of the network are more or less known. In this context, the data is primarily used for checking predictions based on such models and not necessarily for reconstructing the networks themselves. Similar conclusions can be made about various stochastic approaches to simulating behavior of such networks. However, such stochastic generalizations are likely to offer a more realistic result in biochemical networks involving molecules of very low abundance, as is the case in gene

expression regulation. In contrast to these quantitative models, the Boolean network model offers a relatively straightforward approach to reconstructing the topology of the network based on discretized data (e.g. the expression data for each gene at each experiment is reduced to two states expressed/not expressed). However, it has been argued that the binary 0/1 representation of network components is inadequate in many situations.

(Insert Figure 7 here)

Probabilistic graphical models, Bayesian Networks in particular, seem to be capable of capturing the rich topological structure, integrating components operating on various scales and representing stochasticity of both underlying biologic processes and the noise inherent in the data. In this statistical approach, the behavior of the network is expressed as the joint probability distribution of measurements that can be made on elements of the network. The structure of the network is described in terms of the Directed Acyclic Graph (DAG) [91]. Nodes in the network correspond to the elements of the network and directed edges specify the dependences between the components. DAG specifies the dependence structure of the network through the Markov assumption that the node is statistically independent of its non-descendants given its parents. Well-established inferential procedures allow for the data-driven reconstruction of the network topology and specific interactions along with the corresponding measures of confidence in the estimated structure and model parameters [92].

The hypothetical Bayesian network in Figure 7A describes the interaction of four different genes. Assuming that G1, G2, G3 and G4 are variables the describing level of expression of these genes, one of the probabilistic statements encoded by the topology of

this network is that the expression level of Gene4 is independent of expression levels of Gene1 and Gene2 given the expression level of Gene3. In other words, while the expression levels of Gene1 and Gene2 do affect the expression level of Gene4, they do it only through their effect on Gene3, which in turn affects the expression level of the Gene4. By just looking only at the expression of these four genes across different experiments, they would all appear to be correlated to various degrees. The goal of the analysis could then be to infer the most likely network topology explaining these correlations. For example, the network in Figure 7B would induce a similar pattern of correlations between these 4 genes. However, effects of Gene1 and Gene2 on Gene4 is direct and not through their regulation of Gene 3. Given a sufficient amount of data, one can distinguish between these two structures and establish the most likely topology describing their interactions which could then be tested experimentally.

The ability to incorporate various levels of prior knowledge through informative priors about the structure and the local probability distributions [93] allows Bayesian networks to potentially serve as the model of choice for encoding the current knowledge and the analysis of new data on the background of the current knowledge. Predictions about the future behavior of the network can take into account all sources of uncertainties: uncertainty about the estimated parameters of the networks, structure of the network and the stochastic nature of the biologic system modeled by the network.

Reference List

1. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E. L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat.Biotechnol. 14, 1675-1680.

2. Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., Linsley, P. S. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat.Biotechnol. 19, 342-347.

3. DeRisi, J. L., Iyer, V. R., Brown, P. O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680-686.

4. Kerr, K. M.,Churchill, G. A. 2001. Experimental design for gene expression microarrays. Biostatistics 2, 183-201.

5. Colantuoni, C., Henry, G., Zeger, S., Pevsner, J. 2002. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. Biotechniques 32, 1316-1320.

6. Hill, A. A., Brown, E. L., Whitley, M. Z., Tucker-Kellogg, G., Hunter, C. P., Slonim, D. K. 2001. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. Genome Biol. 2, RESEARCH0055.

7. Yang, Y., Dudoit, S., Luu, P., Speed, T. 2000. Normalization for cDNA microarray data. SPIE BiOS 2001, San Jose, California.

8. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W., Albertson, D. G. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat.Genet. 20, 207-211.

9. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., Brown, P. O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat.Genet. 23, 41-46.

10. Forozan, F., Mahlamaki, E. H., Monni, O., Chen, Y., Veldman, R., Jiang, Y., Gooden, G. C., Ethier, S. P., Kallioniemi, A., Kallioniemi, O. P. 2000. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. Cancer Res. 60, 4519-4525.

11. Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A. L., Brown, P. O. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the

transcriptional program of human breast tumors. Proc.Natl.Acad.Sci.U.S.A 99, 12963-12968.

12. Fejzo, M. S., Godfrey, T., Chen, C., Waldman, F., Gray, J. W. 1998. Molecular cytogenetic analysis of consistent abnormalities at 8q12-q22 in breast cancer. Genes Chromosomes.Cancer 22, 105-113.

13. Nadon, R.,Shoemaker, J. 2002. Statistical issues with microarrays: processing and analysis. Trends Genet. 18, 265-271.

14. Cleveland, W. S.,Devlin, S. J. 1988. Locally-weighted Regression: An Approach to Regression Analysis by Local Fitting. J.Am.Statist.Assoc. 83, 596-610.

15. Dudoit, S., Yang, Y., M.J., Speed, T. P. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12, 111-139.

16. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., Speed, T. P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30, e15.

17. Bolstad, B. M., Irizarry, R. A., Astrand, M., Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19, 185-193.

18. Fielden, M. R., Halgren, R. G., Fong, C. J., Staub, C., Johnson, L., Chou, K., Zacharewski, T. R. 2002. Gestational and lactational exposure of male mice to diethylstilbestrol causes long-term effects on the testis, sperm fertilizing ability in vitro, and testicular gene expression. Endocrinology 143, 3044-3059.

19. Ma, X. J., Salunga, R., Tuggle, J. T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B. M., Zhou, Y. X., Varnholt, H., Smith, B., Gadd, M., Chatfield, E., Kessler, J., Baer, T. M., Erlander, M. G., Sgroi, D. C. 2003. Gene expression profiles of human breast cancer progression. Proc.Natl.Acad.Sci.U.S.A.

20. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de, R. M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein, L. P., Borresen-Dale, A. L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc.Natl.Acad.Sci.U.S.A 98, 10869-10874.

21. Churchill, G. A. 2002. Fundamentals of experimental design for cDNA microarrays. Nat.Genet. 32 Suppl, 490-495.

22. Kerr, K. M., Martin, M., Churchill, G. A. 2000. Analysis of variance for gene expression microarray data. Journal of Computational Biology 7, 819-837.

23. Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R. S. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J.Comput.Biol. 8, 625-637.

24. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., Davis, R. W. 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc.Natl.Acad.Sci.U.S.A 93, 10614-10619.

25. Claverie, J. M. 1999. Computational methods for the identification of differential and coordinated gene expression. Hum.Mol.Genet. 8, 1821-1832.

26. Cui, X.,Churchill, G. A. 2003. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 4, 210.

27. Ideker, T., Thorsson, V., Siegel, A. F., Hood, L. E. 2000. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J.Comput.Biol. 7, 805-817.

28. Efron, B., Tibshirani, R., Storey, J. D., Tusher, V. 2001. Empirical Bayes Analysis of a Microarray Experiment. JASA 96, 1151-1160.

29. Baldi, P.,Long, A. D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics. 17, 509-519.

30. Tusher, V. G., Tibshirani, R., Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc.Natl.Acad.Sci.U.S.A 98, 5116-5121.

31. Benjamini, Y.,Hochberg, Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society B **57**, 289-300.

32. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. Mol.Biol.Cell 11, 4241-4257.

33. Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W. H., Tom, E., Mack, D. H., Levine, A. J. 2000. Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. Genes Dev. 14, 981-993.

34. Perou, C. M., Jeffrey, S. S., van de, R. M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O., Botstein, D. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc.Natl.Acad.Sci.U.S.A 96, 9212-9217.

35. 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., . 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-511.

36. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc.Natl.Acad.Sci.U.S.A 96, 6745-6750.

37. 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. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

38. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., Davis, R. W. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol.Cell 2, 65-73.

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

40. Everitt, B. S. 1993. Cluster Analysis. Edward Arnold, London.

41. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., Church, G. M. 1999. Systematic determination of genetic network architecture. Nat.Genet. 22, 281-285.

42. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., Golub, T. R. 1999. Interpreting patterns of gene expression with

self-organizing maps: methods and application to hematopoietic differentiation. Proc.Natl.Acad.Sci.U.S.A 96, 2907-2912.

43. McLachlan, J. G. and E. K. Basford. 1987. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

44. Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., Ruzzo, W. L. 2001. Model-based clustering and data transformations for gene expression data. Bioinformatics. 17, 977-987.

45. McLachlan, G. J., Bean, R. W., Peel, D. 2002. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18, 413-422.

46. Medvedovic, M.,Sivaganesan, S. 2002. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics 18, 1194-1206.

47. Yeung, K. Y., Medvedovic, M., Bumgarner, R. E. 2003. Clustering Gene Expression Data with Repeated Measurements. Genome Biology 4, R34.

48. Cheeseman, P. and J. Stutz. 1996. Bayesian Classification (AutoClass): Theory and Results, p. 153-180. *In*: Advances in Knowledge Discovery and Data Mining.

49. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., Meyerson, M. 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc.Natl.Acad.Sci.U.S.A 98, 13790-13795.

50. Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S., Kato, K. 2003. Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. Genome Biol. 4, R21.

51. 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. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

52. Thomas, R. S., Rank, D. R., Penn, S. G., Zastrow, G. M., Hayes, K. R., Pande, K., Glover, E., Silander, T., Craven, M. W., Reddy, J. K., Jovanovich, S. B.,

Bradfield, C. A. 2001. Identification of toxicologically predictive gene sets using cDNA microarrays. Mol.Pharmacol. 60, 1189-1194.

53. Waring, J. F., Jolly, R. A., Ciurlionis, R., Lum, P. Y., Praestgaard, J. T., Morfitt, D. C., Buratto, B., Roberts, C., Schadt, E., Ulrich, R. G. 2001. Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. Toxicol.Appl.Pharmacol. 175, 28-42.

54. Waring, J. F., Ciurlionis, R., Jolly, R. A., Heindel, M., Ulrich, R. G. 2001. Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. Toxicol.Lett. 120, 359-368.

55. Li, L., Weinberg, C. R., Darden, T. A., Pedersen, L. G. 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics. 17, 1131-1142.

56. Cho, J. H., Lee, D., Park, J. H., Kim, K., Lee, I. B. 2002. Optimal approach for classification of acute leukemia subtypes based on gene expression data. Biotechnol.Prog. 18, 847-854.

57. Alizadeh, A. A., Ross, D. T., Perou, C. M., van de, R. M. 2001. Towards a novel classification of human malignancies based on gene expression patterns. J.Pathol. 195, 41-52.

58. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J. R., Nevins, J. R. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc.Natl.Acad.Sci.U.S.A 98, 11462-11467.

59. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., Meltzer, P. S. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat.Med. 7, 673-679.

60. Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: Data mining, inference, and prediction. Springer-Verlag, New York.

61. Webb, A. 1999. Statistical Pattern Recognition. Oxford University Press Inc., New York.

62. Dudoit, S., Fridlyand, J., Speed, T. P. 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. JASA 97, 77.

63. 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. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537.

64. Nguyen, D. V.,Rocke, D. M. 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18, 39-50.

65. Zien, A., Kuffner, R., Zimmer, R., Lengauer, T. 2000. Analysis of gene expression data with pathway scores. Proc.Int.Conf.Intell.Syst.Mol.Biol. 8, 407-417.

66. Boulton, S. J., Gartner, A., Reboul, J., Vaglio, P., Dyson, N., Hill, D. E., Vidal, M. 2002. Combined functional genomic maps of the C. elegans DNA damage response. Science 295, 127-131.

67. Ge, H., Liu, Z., Church, G. M., Vidal, M. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat.Genet. 29, 482-486.

68. Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., Vidal, M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res. 11, 2120-2126.

69. Vidal, M. 2001. A biological atlas of functional maps. Cell 104, 333-339.

70. Holmes, I.,Bruno, W. J. 2000. Finding regulatory elements using joint likelihoods for sequence and expression profile data. Proc.Int.Conf.Intell.Syst.Mol.Biol. 8, 202-210.

71. Barash, Y.,Friedman, N. 2002. Context-specific bayesian clustering for gene expression data. J Comput Biol. 9, 169-191.

72. Segal, E., Yelensky, R., Koller, D. 2003. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19 Suppl 1, I273-I282.

73. Segal, E., Wang, H., Koller, D. 2003. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19 Suppl 1, I264-I272.

74. Ideker, T., Galitski, T., Hood, L. 2001. A new approach to decoding life: systems biology. Annu.Rev.Genomics Hum.Genet. 2, 343-372.

75. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292, 929-934.

76. Fritz, B., Schubert, F., Wrobel, G., Schwaenen, C., Wessendorf, S., Nessling, M., Korz, C., Rieker, R. J., Montgomery, K., Kucherlapati, R., Mechtersheimer, G., Eils, R., Joos, S., Lichter, P. 2002. Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. Cancer Res. 62, 2993-2998.

77. Weiss, M. M., Kuipers, E. J., Postma, C., Snijders, A. M., Siccama, I., Pinkel, D., Westerga, J., Meuwissen, S. G., Albertson, D. G., Meijer, G. A. 2003. Genomic profiling of gastric cancer predicts lymph node status and survival. Oncogene 22, 1872-1879.

78. Niehrs, C.,Meinhardt, H. 2002. Modular feedback. Nature 417, 35-36.

79. Ideker, T. E., Thorsson, V., Karp, R. M. 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. Pac.Symp.Biocomput.305-316.

80. Voit, E. O. 2002. Metabolic modeling: a tool of drug discovery in the post-genomic era. Drug Discov.Today 7, 621-628.

81. Voit, E. O.,Radivoyevitch, T. 2000. Biochemical systems analysis of genome-wide expression data. Bioinformatics. 16, 1023-1037.

82. Gillespie, D. T. 1977. Exact stochastic simulation of cupled chemical reactions. J.Phys.Chem. 81, 2340-2361.

83. Kierzek, A. M. 2002. STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm. Bioinformatics 18, 470-481.

84. McAdams, H. H.,Arkin, A. 1999. It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 15, 65-69.

85. McAdams, H. H.,Arkin, A. 1998. Simulation of prokaryotic genetic circuits. Annu.Rev.Biophys.Biomol.Struct. 27, 199-224.

86. McAdams, H. H.,Arkin, A. 1997. Stochastic mechanisms in gene expression. Proc.Natl.Acad.Sci.U.S.A 94, 814-819.

87. D'haeseleer, P., Liang, S., Somogyi, R. 2000. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 16, 707-726.

88. Friedman, N., Linial, M., Nachman, I., Pe'er, D. 2000. Using Bayesian networks to analyze expression data. J.Comput.Biol. 7, 601-620.

89. Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. Science 303, 799-805.

90. de Jong, H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. J.Comput.Biol. 9, 67-103.

91. Cowell, R. G., P. A. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. 1999. Probabilistic Networks and Expert Systems. Springer, New York.

92. Heckerman, D. 1998. A tutorial on learning with Bayesian networks, p. 301-354. *In*: M. I. Jordan (ed.), Learning in graphical models. Kluwer Academic Publishers, Dordrecht, The Netherlands.

93. Heckerman, D., Geiger, D., Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20, 197-243.

Figure 1. Scatter plot of log-ratios vs average log-intensity in a typical microarray experiment. The line represents the local regression line used for centering log-ratios.

Figure 2. Clustering cell-cycle expression data A: two individual gene expression profile alone; B: the same two profiles in the background of two major underlying expression patterns C,D: clusters defining underlying patterns of expression.

Figure 3. Hierarchical clustering of the cell cycle genes from Figure 2. Each line in the color-coded display corresponds to the expression profile of one gene with the red color denoting high expression and green color denoting low expession.

Figure 4. Right: distribution of between clusters and within clusters distances based on pair-wise correlations. Left: distribution of between- and within-clusters distances based on posterior probabilities of expression calculated from the Bayesian Infinite Mixture Model

Figure 5. Classifying hypothetical tumor tissues based on the expression levels of two genes. A) Scatter plot of data for 100 samples. B) Underlying probability distribution of expression data for two classes. C) Separating two classes based on Gene1 data only. D) Separating two classes based on Gene 2 data only. E) Optimal linear classifier for the two classes based on the linear combination of Gene1 and Gene2 expression data.

Figure 6. Principal component based classifier

Figure 7. Two alternative Bayesian networks explaining the correlation structure between expression measurements of four hypothetical genes.

Figure 1

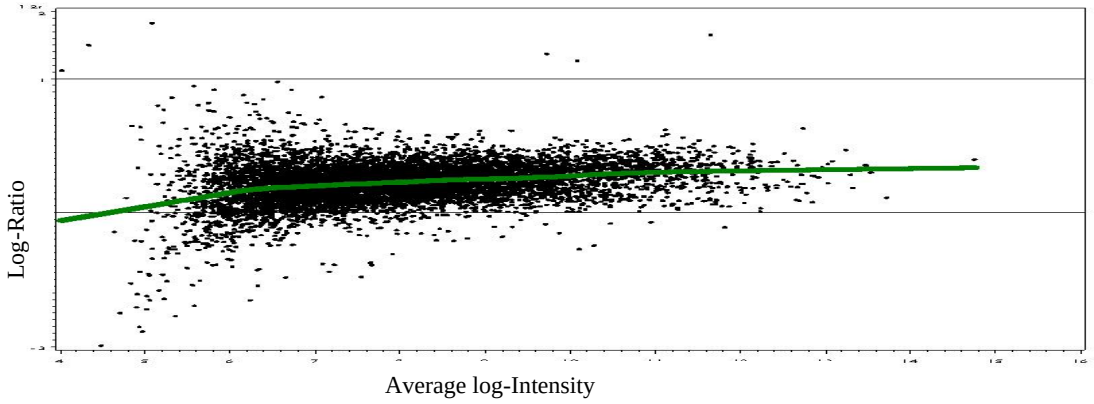


Figure 2

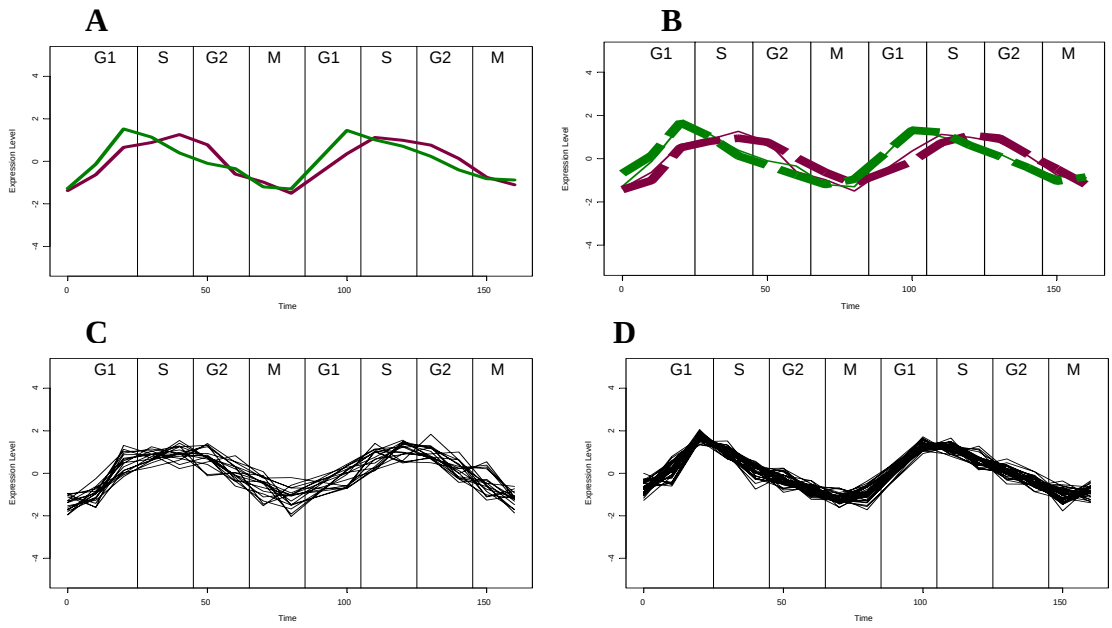


Figure 3

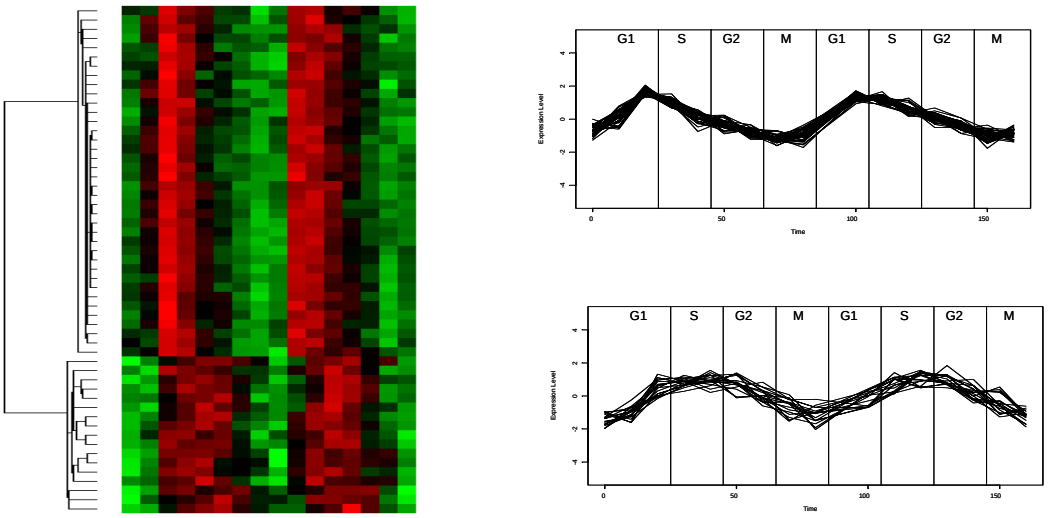


Figure 4

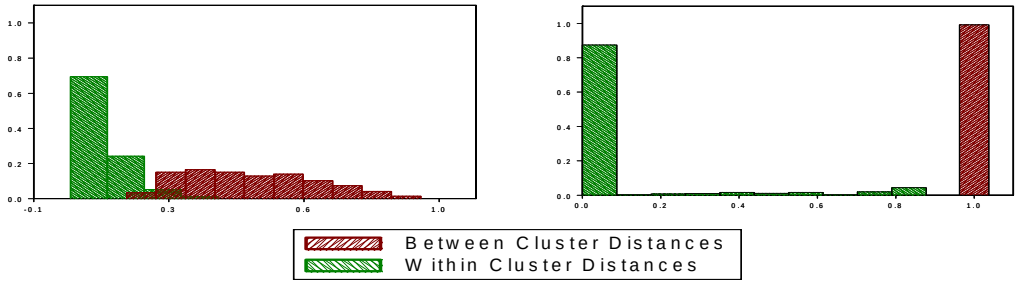

1.0
0.8
0.6
0.4
0.2
0.0
-0.1
0.3
0.6
1.0
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Between Cluster Distances
Within Cluster Distances

Figure 5

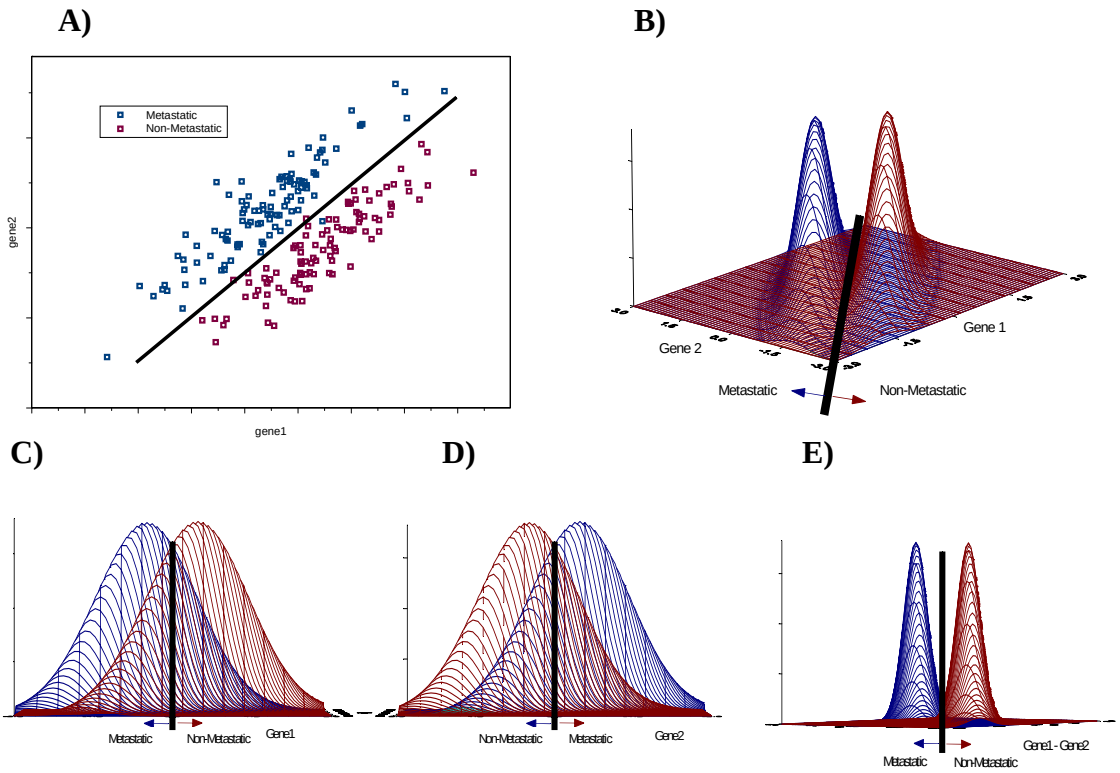

A)
B)
Metastatic
Non-Metastatic
gene2
gene1
Gene 1
Gene 2
Metastatic
Non-Metastatic
C)
D)
E)
Metastatic
Non-Metastatic
Gene1
Non-Metastatic
Metastatic
Gene2
Metastatic
Non-Metastatic
Gene1 - Gene2

Figure 6

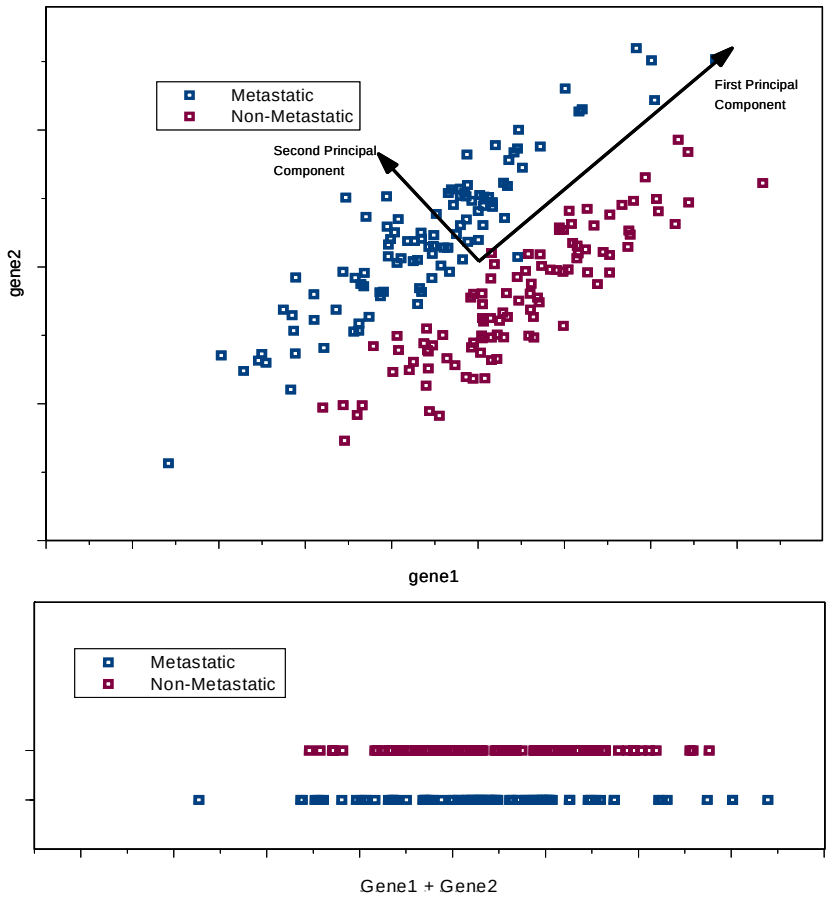
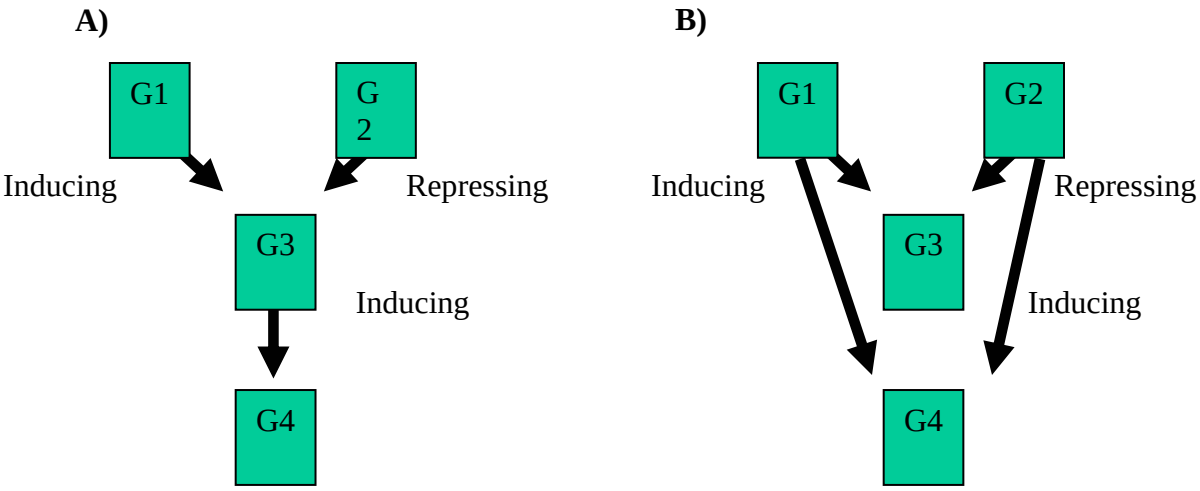

Metastatic
Non-Metastatic
First Principal Component
Second Principal Component
gene2
gene1
Metastatic
Non-Metastatic
Gene1 + Gene2

Figure 7


A)
G1
G
2
Inducing
Repressing
G3
Inducing
G4
B)
G1
G2
Inducing
Repressing
G3
Inducing
G4