Published in final edited form as:
Annu Rev Biomed Eng. 2007 ; 9: 205–228. doi:10.1146/annurev.bioeng.9.060906.151904.

Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities

I.P. Androulakis[1], **E. Yang**[1], and **R.R. Almon**[2]

[1]Biomedical Engineering Department, Rutgers University, Piscataway, New Jersey 08854

[2]Department of Biological Sciences, and Department of Pharmaceutical Sciences, State University of New York at Buffalo, Buffalo, New York 14260; almon@eng.buffalo.edu

Abstract

Monitoring the change in expression patterns over time provides the distinct possibility of unraveling the mechanistic drivers characterizing cellular responses. Gene arrays measuring the level of mRNA expression of thousands of genes simultaneously provide a method of high-throughput data collection necessary for obtaining the scope of data required for understanding the complexities of living organisms. Unraveling the coherent complex structures of transcriptional dynamics is the goal of a large family of computational methods aiming at upgrading the information content of time-course gene expression data. In this review, we summarize the qualitative characteristics of these approaches, discuss the main challenges that this type of complex data present, and, finally, explore the opportunities in the context of developing mechanistic models of cellular response.

Keywords

microarrays; bioinformatics; regulation; clustering; pharmacogenomics

TEMPORAL GENE EXPRESSION ANALYSIS

At any given time a cell will only express a small fraction of the thousands of genes in the organism's genome. Expressed genes reflect the structure and functional capacities of the cell as well as the ability of the cell to respond to external stimuli. In a complex organism, external stimuli to a great extent take the form of chemical messages whose purpose is to coordinate the function of the complex society of cells (1). Gene arrays, which measure the level of mRNA expression of thousands of genes simultaneously, provide a method of high-throughput data collection necessary for obtaining the scope of data required for understanding the complexities of living organisms. Monitoring the change in expression patterns over time using gene arrays provides an approach for capturing the multidimensional dynamics of complex biological systems. By using gene arrays in a time series paradigm, we are able to observe the emergence of coherent temporal responses of many interacting components. The data should provide the basis for understanding evolving

yannis@rci.rutgers.edu, eyang@eden.rutgers.edu

but complex biological processes, such as disease progression, growth, development, and drug responses.

Global gene expression analysis has been celebrated as a major revolution in modern biology (2). The ability to monitor simultaneously the expression of the genes composing the entire genome has generated unimaginable possibilities (3–5). Despite some criticism regarding the cross-platform reproducibility of expression experiments (6, 7), more recent evidence (8, 9) supports the informative nature of the experiment and the importance of the approach (10). Microarray analysis has found widespread applications from characterizing terminal states, i.e., benign versus malignant tumors (11), to attempts to decipher the evolution of complex diseases and cell fates (12– 16). Hence, the nature of the data broadly defines the nature of the problems to be addressed.

Boundary problems use the expression measurements as feature vectors that characterize static points in multidimensional spaces. Therefore, multiple samples, for example, from the same tissue of different patients (diseased/nondiseased) would define a database of multidimensional feature objects with as many dimensions as genes whose mRNA has been quantified and as many objects as the number of patients monitored. Critical questions then arise, such as how to identify coherent patterns, i.e., combinations of up- or downregulated genes that distinctly characterize the two or more classes of patients (17, 18).

Monitoring the change in expression patterns over time provides a profoundly different type of information. Instead of concentrating at terminal points in time of binary nature (benign versus malignant, type A versus type B, etc.), we now have the opportunity to observe the emergence of coherent temporal responses of many interacting components. The orchestrated response of an organism to an external stimulus and the monitoring of the temporal progression offer numerous opportunities for reverse-engineering the mechanisms that regulate the host responses (19). The latter, in turn, will define the rational foundation for the generation of testable hypotheses. Thus, the challenge becomes how to upgrade the information content of such multidimensional trajectories to address critical questions such as the characterization of the state of evolution of a system; the identification of activated pathways, their relation, and the rate limiting steps; and the synthesis of interaction networks and the characterization of points of control.

A number of questions could potentially be addressed that fall, broadly, under the following categories (20):

1. Biological systems analysis: Specific systems are monitored over time and information is assembled to understand the driving dynamics. Prototypical examples include cell cycles and circadian clocks (21, 22).
2. Response dynamics: Systems are subjected to controlled perturbations and the broad gene expression response of the system is monitored over time. Examples include drug dosing and defined trauma (39, 50).
3. Development: Morphing of organisms during development involves complex sequences of cell proliferation and differentiation. Many models have been used

over the years (23) to address the process of development. Particularly exciting are the opportunities offered by recent advances in stem cell differentiation (24).

4. Disease progression: Genome-wide temporal profiling offers the possibility of elucidating the underlying pathophysiologies of human diseases (25). Instead of focusing on predefined hypotheses, global expression offers the possibility of unraveling the systemic evolution of pathological conditions.

The purpose of this review is not to provide a detailed account of the enormous complexities and uncertainties surrounding the collection of the required data (14, 26). The following sections will highlight conceptually the basic foundations of the computational approaches that have been recently proposed for the analysis of temporal gene expression data, the opportunities that exist, and, more importantly, the challenges that need to be addressed.

METHODS

The goal of temporal gene expression analysis is to identify broad sequences of molecular events in time. Such sequences can be associated with an ongoing biological process, such as the cell cycle, circadian rhythms, or development, or can be initiated by some input perturbation, such as the administration of a drug or a defined trauma. The host state is defined as the ensemble of all possible metabolites, proteins, small molecules, etc., which define the observed phenotype of the organism. For some purposes, such as the consideration of circadian rhythms, the host state may be dynamic. Any perturbation sets in motion the information transfer that defines the blueprint for the production of the relevant components of the response by activating appropriate genes whose transcription to mRNA and subsequent translation to proteins catalyzes critical functions. Therefore, the implicit underlying assumption of transcription profiling is that gene expression is causal to phenotypic responses through production of specific proteins coded by the expressed mRNAs.

One of the major limitations of monitoring exclusively mRNA transcripts is the role of posttranslation modifications, mRNA stability, and other destabilizing and complicating factors that render the products of transcription (mRNA) an inaccurate proxy for the abundance of active products of translation (27, 28). Nevertheless, analysis of the products of transcription has already provided significant insight and is undoubtedly a critical source of information. Associated with expression profiling is the implicit assumption that gene expression is tightly controlled by a fine-tuned, intricate, and robust regulatory mechanism that appropriately activates and deactivates the machinery guiding the expression of genes. By now it is almost taken for granted that genes exhibiting similar responses to signals ought to be controlled by similar regulatory mechanisms. This is often referred to as the guilt by association principle (29). Therefore, identifying coherent expression responses is important in the sense that if coexpression can be linked to coregulation then the underlying machinery driving expression can be isolated to smaller groups, deciphered, and quantified. One of the most critical problems is to verify the common regulatory mechanism (30). Hence, the first and most critical step in this endeavor it to identify those measured transcripts that appear to be somehow correlated to each other. From a computational point of view, this problem

belongs to a more general class, namely, the characterization (indexing and clustering) of multidimensional trajectories (31).

Numerous methods have been developed and applied to this very challenging problem in the context of analyzing gene expression data (20, 32, 33). At the core of all the methods is the concept of similarity and we will segregate the approaches based on the relative use of this term. Given that we treat expression measurements as multidimensional trajectories, sampled at discrete points, the first definition of similarity ought to be based on some kind of point-wise metric measuring the distance among the various objects, using the relative mRNA abundance as the multidimensional feature set. Various metrics have been utilized, most notably Euclidean distances. Methods such as k-means and hierarchical clustering by and large fall under this family of approaches. Pair-wise comparisons are made and then combined to assess the relative degree of similarity. A second family of methods assumes the existence of a finite set of undetermined processes that generate the observations. Two trajectories are thus declared similar if they are the product of similar processes. Therefore, the comparison is made not in the finite-dimensional space of raw data, but in the infinite-dimensional space of the functions that generate this data. Finally, a third general family of methods defines the similarity in terms of global features and characteristics of the trajectories. Instead of focusing on point-wise differences or specific functional forms giving rise to data, these methods aim at identifying structural characteristics of the responses and define similarity based on pattern recognition methods that aim at finding salient changes and similarities in the responses.

A comprehensive review on clustering methods was recently presented in Reference 34. In the following section we attempt to partition qualitatively the methods by describing the essential characteristics of each approach.

Point-Wise Distance-Based Clustering Methods

Recently, a very nice and concise review of distance-based clustering methods was presented (35). Without loss of generality, we assume that the data are given in a matrix form:

$$\mathrm{E}=\{\mathrm{E}(\mathrm{i},\mathrm{t}\},\quad \mathrm{i}=1,\ldots,\mathrm{N_g},\quad \mathrm{t}=1,\ldots,\mathrm{N_t},$$

assuming the Ng genes are measured at Nt discrete time points. The goal of point-wise distance-based clustering methods (PwDbM) is to quantify the distance between any two samples and agglomerate samples that fall within a predefined threshold. Usual metrics include various definitions of norm-based distances and combinations of known correlation expressions. Indicative definitions are provided in **Table 1**.

The fundamental difference between the various distance-based methods is in the way these distances are being combined to identify the proper partitioning of the data. Two major classes of methods exist: (*a*) partitioning and (*b*) hierarchical. Among partitioning methods, the prototypical example is k-means clustering, although self-organizing maps (SOM) also follow the same basic principles (15). Using these methods, a predetermined number of

partitions of the feature space that are defined by a nominal center are constructed. Points are subsequently assigned to each partition based on their relative proximity to the center with the overall objective of minimizing the distance of each point from its respective center:

$$\text{Error}=\sum_{k=1}^{N_k}\sum_{i\in C_k}\sum_{t=1}^{N_t}|E(i,t)-M(k,t)|^2.$$

In this general definition of the objective (Error) the objects (temporal gene expression of Ng genes) have been partitioned to Nk clusters. The purpose of the optimization search is to assign each profile to one of these clusters, such that the sum of the distances of each profile from the center of the cluster it has been assigned to, M(k, t), is minimized.

Recently, interesting combinations of k-means and kernel methods have emerged (36, 37). Essentially, kernel methods aim at identifying appropriate nonlinear transformations of the original data through the use of kernel functions that render the data linearly separable (38). The distances are then defined on the transformed data rather than the original. Even though kernel-based methods have a number of advantages, such as creating separability through transformations or relative robustness to noise, they do require the identification of a number of input parameters that would render the estimation problem user-specific, and the appropriate estimation of the necessary parameters is not trivial.

Hierarchical methods create a hierarchy of relative distances (hence the name) and place multidimensional points along a one-dimensional axis based on the relative distance between points. The result of the analysis is presented in the form of a dendogram in which the relative positions of points defines their relative distance as well. The dendogram is essentially a binary tree with the root representing the entire data set and each leaf node representing a data object. Intermediate nodes represent the extent to which objects are close to each other. Among the strongest criticisms raised for classical hierarchical clustering is the fact that these algorithms are lacking robustness to noise and are therefore sensitive to outliers.

The literature in terms of applications of distance-based clustering in the analysis of microarray data is really abundant. Typical examples of the applications of such methods include the work of Eisen and collegues and Gash and collegues (39, 40) for hierarchical clustering and the work of Tavazoie and collegues (41) using k-means clustering. Eisen et al. (39) clustered expression data in the budding yeast *Saccharomyces cerevisiae* to deduce that clustering gene expression data grouped together genes of known similar function, interpreting this observation as an indication of the status of cellular processes. They applied hierarchical clustering where the linkage was determined using a similarity score based on a correlation coefficient. Gash et al. (40) evaluated the response of yeast to numerous environmental perturbations to unravel the effects of environmental stresses on the cell. Tavazoie et al. (41) measured 15 time points across two cell cycles of *Saccharomyces cerevisiae* and analyzed the results using variance normalized profiles and k-means

clustering. Clusters were subsequently characterized based on their relative functional enrichment to demonstrate the concentration of similar functions within each cluster.

The use of correlation-based distance metrics, such as Pearson's coefficient, represents a variation upon Euclidean distance by providing a scale-free distance metric between two feature vectors. This metric has been widely used as well and is particularly useful when the baseline magnitude of different mRNA messages differ greatly.

Model-Based Clustering Methods

Model-based clustering methods (MbCMs) (42–46) shift the similarity emphasis from the data to an unknown model that describes the data. These are methods based on variants of mixture-models (44). The general idea is that finite mixtures of distributions provide a flexible approach to modeling. Therefore, each point (i.e., expression profile) is taken to be the outcome of the superposition of a finite number of processes, much like expansion over a basis set, with a number of unknown parameters to be determined based on the available experimental data. Therefore, the objective is to identify this underlying set of functions (models) whose appropriate combination (mixture) assigns the data properly. The existence of such a finite and coherent set of basis functions indicates the existence of an underlying set of limited common processes that give rise to the observed behavior. Without loss of generality we will use the formalism of Pan et al. (42) to illustrate the approach. Let y denote any measurement, then each such data point is assumed to be the superposition of distributions given by

$$f(y;f_g) = \sum_{i=1}^{g} \pi_i \varphi(y;\mu_i, V_i).$$

The density function, φ, depends on appropriate parameters, μ and V, which, for example, could correspond to the mean and covariance matrix. These mixtures are appropriately weighted by means of the mixing proportion π. Thus the model to be estimated based on the data is composed of the triplet (π, μ, V), with parameters determined through the use of appropriate expectation-maximization algorithms.

It is important to realize the fundamental difference between the distance-based and the model-based approaches. The emphasis is now placed on the speculated underlying model, thus making the approach more robust in the presence of noisy data. However, the assumption is that such underlying processes do exist, requiring that the data follow a set of predetermined distributions. A slight variation was recently proposed in Reference 45, whereby an autoregressive model able to account for time delays was assumed to exist and was subsequently estimated based on the data. Similar in spirit are methods based on hidden Markov models (47–49), which assume an underlying HMM describing the sequence of events corresponding to the transformed temporal gene expression profiles. An interesting method was proposed (50) in which a linear dynamic model is invoked to simulate the level of mRNA that gives rise to time-dependent profiles, which are considered to be sums of exponentials. The associated parameters of the model are estimated through nonlinear regression. The number of exponentials is also minimized by making use of the concept of

information theoretic arguments quantifying Occam's Razor, such as minimum description length (51) and Akaike information criterion (52). Model-based approaches have been proposed (53, 54) that consider the cell to be a system where the behaviors (responses) of the cell depend completely on the current internal state plus any external inputs, and the proposed method regards a time-course gene expression dataset as a set of time series generated by a number of stochastic processes. Each stochastic process defines a cluster and is described by an autoregressive model. Along those lines, significance analysis of time-course microarray experiments was also recently proposed as a competitive alternative (55). The method is applicable to detecting changes in expression over time within a single biological group and to detecting differences in the behavior of expression over time between two or more groups.

In summary, the fundamental assumption of model-based approaches is that the expression profiles are clusters in the space of the functionals that characterize them. The question thus becomes how to identify this functional decomposition of the data, as opposed to decomposing the raw data. One of the key drivers for such methods is the speculation that gene expression profiles are generated by time-dependent models, in the sense that the current state is a function of the cellular state at previous times (45). Therefore, these methods attempt to quantify this assumption.

Feature-Based Clustering Methods

Feature-based clustering methods (FbCMs) aim at detecting salient features and local or global shapes characteristic of the expression profiles. One of the key motivating arguments for such methods is the realization that in the presence of noise and uncertainties associated with measuring mRNA abundance, looking for specific quantifiable metrics may not necessarily yield the most informative interpretation. Instead, robust, coherent, and dominating qualitative features and similarities could be a more informative proxy for the information content of the expression experiment. The raw data are transformed to sequences of events or symbols, and these are further analyzed for consistencies, either local or global (56). Looking for general shapes as opposed to quantifying distances allows for, among other things, a more flexible representation, which uncovers more intricate relations among expression profiles, such as time shifts and inversion in expression profiles (57). Syeda-Mahmood (58) has proposed a pattern recognition approach aimed at capturing salient features of the time-varying gene expression patterns, such as inflection points based on the idea that dissimilar curves, when represented as two-dimensional curves, show a significant number of twists and turns. A new framework was recently proposed (59, 60). Both approaches share a critical similarity: The transformation of the raw expression data to a sequence of symbols and the subsequent analysis of the symbolic representation of the time series. This type of approach, motivated by recent advances in the symbolic representation of streaming data (61), effectively reduces the dimensionality of the time series from an infinite-dimensional space (continuous representation of expression level) to a finite, quantized representation where each profile is represented by a sequence of symbols. In effect, the most significant variation introduced by these methods is a fine-grained clustering, with a potentially enormous number of clusters defined.

There have been subsequent significant variations in both methods. One is based on the relative probabilities of each symbolic sequence (59) and the other is based on the ability of selected subsets to reproduce the overall dynamic response (60), with selection criteria ranking the importance of the respective clusters. Because the method proposed by Ernst & Bar-Joseph (59) needs to postulate a priori the putative sequence of events, the method is best suited for short time series, whereas the method proposed Yang et al. (60) has complexity that is effectively linear with respect to the number of genes.

Interesting algorithms for clustering (expression) data are emerging, exploring graph-theoretic properties. We discuss them in Feature-Based Clustering Methods because essentially the structure of the graph created from the original data is analyzed. In other words, in graph-based methods, the nature, structure, properties, and characteristics of a graph whose edges represent data points and the arcs relative distances between those points are treated as the features to be further analyzed. Thus, subgraphs are formed and identified containing enough nodes for effective similarity computations. Effectively, a tree representation converts the multidimensional problem to a tree partitioning problem.

Among the most popular methods are those that explore the concept of the minimum spanning tree (62) and effectively attempt to identify cliques within the data set (63, 64). An interesting extension of the MST concept is discussed in Reference 65 where a metric for assessing the clustering potential based on geometric arguments is presented. Assessing the “clusterability” potential of a dataset a priori will greatly enable further analyses. In the case of temporal data, this remains an open question.

FbCMs offer a higher degree of flexibility. Appropriate selection of characteristic features offers the possibility of defining a time-course representation using variables that potentially capture intrinsic and implicit characteristics of the responses. Undoubtedly, the definition of such features results in some kind of lumping of the response, which can potentially result in loss of fine-grain detail.

Clustering Across Conditions

Each gene expression experiment is essentially a set of observations generated from a single perturbation of the system, whether it is a particular growth condition, an injury, or the administration of a drug. It can be argued that extracting information from a single perturbation contains little information. Therefore, increased methods that attempt to simultaneously analyze multiple conditions are continually attracting increased attention (66). Bi-clustering in the context of gene expression analysis was first suggested by Church (67) and refers to simultaneous clustering across “columns” and “rows” in expression data by expanding the concept of similarity so that it does not become a function of pairs of genes or pairs of conditions, as is normally the case, but rather it becomes a measure of coherence of the genes and conditions. Heard and coworkers (68) expanded their original Bayesian model-based agglomerative clustering scheme (69) for time-course data. Their approach uses a spline approximation to capture the temporal variation within each cluster. The approach is particularly intriguing in that the time courses are explicitly treated. Undoubtedly, biclustering (or coclustering) methods hold tremendous promise as more systemic perturbations are becoming available and the need to develop consistent

representations across multiple conditions is required. The underlying assumption, as argued below, is that biological systems are treated as systems in which external perturbations are applied. Therefore, the underlying dynamics should be consistent across conditions independent of the type of the perturbation to assess the biologically informative nature of conclusions drawn from any kind of computational analysis of transcriptional responses.

Summary

Clustering can be characterized as the process of establishing associations. At a conceptual level, the nature of the associations becomes more abstract as the methods evolve from hierarchical, to partition, to model, to feature based. **Figure 1** depicts this increased level of abstraction. Hierarchical clustering would basically associate the two convex and the concave hypothetical patterns of expression by quantifying the relative differences among all members. K-means or SOM will draw an association between the raw profiles and the putative centers of the domains within which each profile lies (centers indicated by the dashed lines). Model-based methods will establish the association between individual profiles and functional representation according to the values of the model parameters. Objects therefore are associated with sets of parameters. Finally, feature-based methods will associate each profile with macroscopic features characteristic of the overall shape of the response. The relative distance between transformed individual members define the proximity. One could argue that even feature-based methods are essentially distance-based methods. However, the transformation from raw data to features relaxes the proximity restrictions and allows for the introduction of soft comparisons.

CHALLENGES

Clustering is by definition an unsupervised task. We can loosely define clustering as the process of organizing objects (expression profiles) into groups whose members are similar in some way. In evaluating the effectiveness of clustering, one could argue that if the groups are similar given some metric, then the clustering was successful. However, the fascinating thing in biology is that similarity in the input space is not the final arbitrator. If the genotype is the input, the actual observable is the phenotype. The information encoded by the objects of the clusters, the mechanisms that brought the objects together, the implications of bringing the objects together, in a nutshell, the biological insight gained by analyzing the objects that were brought together is what will decide the effectiveness of the computational analysis. Therefore, defining the quality of the clustering algorithms is not as straightforward as it may appear. A major challenge in the clustering of microarray data lies in the fact that the metric for evaluating the overall quality of a result is still an open area of research (70). Without a well-defined metric, it becomes difficult to ascertain which method outperforms the others.

Various evaluations have been proposed to quantify the relative advantages of clustering methods for microarray expression data (71, 72), and the metrics for comparison quantified the ability of various methods to generate well-separated clusters. By and large any such comparison is biased and the results to a great extent depend on the specific use of the method as well as the nature and type of data. We believe that a head on comparison between clustering methods based exclusively on some optimality criterion will probably be

misleading. The complexities of the underlying biological system will probably render such analysis mute. Methods should be evaluated, and not compared, based on their ability to generate insight information and it is quite possible that the evaluation could be problem dependent. It is critical to realize that the computational steps, in the context of transcriptional analysis, should be an integrated component of the overall effort and a separate independent activity. Therefore, the effectiveness of a computational approach should be evaluated in the grand scheme of the biological content of a specific analysis.

In the sections that follow, we identify three elements of the computational analysis of time-course gene expression data that we believe could potentially impact the conclusions drawn. Thorough and detailed analyses of the challenges associated with high-dimension clustering in general have been nicely presented elsewhere (73).

Small Sample Size: Information or Noise?

The term “data deluge” is often used in conjunction with microarray data (74). However, this could not be a more misleading characterization. There is no doubt that the observables in a microarray experiment are in the thousands, particularly in temporal experiments. A typical animal study with m replicates (animals) at n time points recording k genes would produce $m \times n \times k$ data points. However, the number of objects, in terms of the machine-learning problem, is quite minimal, and definitely not up to par with the number of features. Examples of the types of objects we refer to include number of patients in a cancer study or number of system perturbations (types/severity of trauma, or drug dosing).

Technological and other practical limitations severely restrict either the number of time points that can be measured or, more importantly, the number of biological and technical replicates that can be used. In the machine-learning community, this is an age-old problem known as learning in almost empty spaces (75). In such cases, it is quite difficult to distinguish noise from structure unless something is known about the underlying concept generating the data. A simple, yet informative example, of errors introduced by subsampling is presented in Reference 76. New technologies that are emerging, such as the living cell array (77), which will provide extensive data at least for model systems, will expose the host system to a wide range of insults and will create a more integrated list of cause-effect relationships. Currently, the only way to condition the data to overcome the lack of a critical mass of observations is to couple the expression data with available prior biological information and analyze simultaneously multiple perturbations. The inability of sparse data to properly capture the complexity of a classification problem is also discussed in Reference 76; however, recent advances in theoretical work on clustering sparse data (78–80) will significantly help.

As noted above, a key complexity of microarray experiments is the essential lack of observables (cell lines or tissue samples) to support the large number of probes monitored. The consequences of the small ratio of features to samples in microar-rays was discussed in Reference 81 and a nice discussion of the impact of the small sample size problem in array expression data is presented in References 82 and 83, which comment on the required optimal number of samples required for robust estimation under certain assumptions regarding the distribution of the measurements. The implication of the ratio of features to

samples is critical, as sparsely populated datasets can very easily lead to random features appearing to be informative. It should be expected that simple minimization of the number of features (genes) in a model need not necessarily provide the best answer. Additional complexity restrictions will have to be imposed to balance the lack of available data, although no definite answer can be provided as no analysis can replace accurate and adequate data. Recently, a novel method for characterizing the information content of short time-course gene expression data was presented by effectively quantifying the random nature of the signal encoded in the expression time series (84).

Knowledge-Based Clustering

Although genome-wide mRNA expression analysis is slowly becoming a routine tool, translating computational results to biological information remains a major challenge. As previously mentioned, one of the key challenges is the improper conditioning of the data. Approaches are being developed that attempt to integrate prior knowledge into the analysis of expression data. In a report by Pan (85), the mixture model for clustering expression data is extended to incorporate gene ontology information as prior knowledge to increase the specificity of the method (85–87). To take advantage of accumulating gene functional annotations, Huang & Pan (88) proposed incorporating known gene functions into a new distance metric that shrinks a gene expression-based distance toward 0 if and only if the two genes share a common gene function.

The incorporation of biological (or any type of prior knowledge) into clustering algorithms will be greatly enabled by recent advances in the area of constraint-based clustering (89, 90) which aims at developing consistent methodologies that incorporate prior knowledge during the analysis, as opposed to postprocessing the results to validate the consistency of the conclusions given what is known about the system. However, one needs to be aware of the constraints that explicit, hard modeling of prior knowledge imposes in terms of discovering new knowledge about the system: Over-restricting and constraining the analysis goes against the very essence of integrative -omics approaches and data-driven (systems) approaches, as opposed to hypothesis-driven research.

Judging the Quality of Gene-Expression Clustering

Clearly there are a number of analytical rationales used to parse genes into groups. However, the quality of all grouping must be judged based on their ability to provide insight into the underlying mechanistic biology. A general concern regarding the validity of existing algorithms stems from the observation that classification algorithms can lead to conflicting results, which are often method dependent (91). The current practice is to evaluate methods based on their ability to generate results consistent with biological reality in terms of functional ontologies and putative transcription factors of coexpressed genes (92–95). Although it is not surprising that different methods yield different results, the fact is, there is a correct answer. Living organisms exist as a fine balance between entropy and enthalpy. Maintaining such a balance requires that the expression of the thousands of genes in the organism's genome be highly coordinated. At any point in time, the amount of mRNA for a particular gene is the balance between its synthesis and degradation. Processes such as circadian rhythms or input perturbations such as drugs can change the amount of an mRNA

by increasing or decreasing synthesis, degradation, or some combination of both. The power of the gene array time series is that it allows the observer to broadly “watch” the dynamics of the system. The objective in conducting time-series experiments is to understand the complex sequence of regulatory events that drives the system. Clustering, regardless of method, attempts to parse genes into groups with certain defined commonalities.

These groups are useful to the biologist to the degree that they represent genes with common mechanisms of regulation. In essence, each proffered group represents a testable hypothesis. If the hypothesis is correct, then certain biological requirements follow. For example, if a group of genes is regulated by a common mechanism, then their response to a different input perturbation should be the same. On the one hand, if the process being examined is natural, such as development, cell cycle, or circadian rhythm, then a perturbation that disrupts the natural process should change the profile over time of all genes in a cluster. To the degree that it does not, then it suggests that the cluster is not entirely valid. On the other hand, if the process being examined is an input to a biological system, such as a drug treatment, then genes that belong in the same cluster should have the same response profile regardless of dosing regimen. In reality, a single temporal response profile probably does not provide sufficient constraint to accomplish biologically valid mechanistic clustering. A second test of the validity of clusters involves the mechanism of control of gene expression. The expression of genes is controlled by transcription factors (TFs). TFs are gene products, proteins that bind to specific sites in the DNA and either promote or inhibit the expression of a gene. Some TFs act on their own or in combination with other TFs, whereas some, such as the glucocorticoid receptor and the estrogen receptor, require the binding of an external ligand for activation. If a group of genes is regulated by a common mechanism, then they should contain common features in their regulatory regions. However, because transcription binding site (TFBS) motifs are short (5–9 base pairs) and fairly degenerate, most putative TFBS matches occur by chance alone and are not functional. One method that has been proposed to identify which TFBS are bona fide functional sites is excluding those that are not in evolutionarily conserved regions. Indeed, the upstream noncoding regions do not evolve in a uniform fashion among sites, but rather show blocks of fairly conserved areas interspersed with fast-evolving stretches. These fast-evolving stretches quickly lose homology with evolutionary time and are subject to insertions and deletions. But nonetheless, even among comparatively distant taxa (e.g., rodents and humans), conserved, alignable segments are preserved (96, 97). Identifying common features in the regulatory region of genes in putative clusters not only provides a degree of validation but also should provide insight into the mechanism of regulation.

OPPORTUNITIES

The traditional way of interpreting time-course expression data is to evaluate the biological similarities implied as a result of expression profile similarities, and currently an enormous number of publications have been presented advocating the potential for such analyses (12, 13, 15, 57, 98).

A healthy body of literature utilizes time-course expression data to reverse-engineer primarily regulatory networks, given that interference at the level of regulation holds

significant promise for drug discovery. Targeting expression by controlling the regulatory process through the corresponding transcription factors is emerging as a viable option for the identification of drug targets (99, 100) and controlling disease progression (101). In recent years, significant efforts have been made experimentally, and computationally, to identify transcription factors, their target genes, and the interaction mechanism that control (regulate) gene expression (102, 103). Prominent examples are the decomposition-based methods, which combine ChiP and microarray data, and inversion of regression techniques to estimate transcription factor activities (TFAs) (104–107). Singular value decomposition and regression methods were combined (19) to reverse-engineer regulatory networks, and in a report by Bussemaker et al. (108), promoter elements were linearly combined to quantify the contribution of the promoter architecture on a gene's expression. Network component analysis (NCA) (109–113) was introduced as an alternative for quantifying the strength of the regulatory interactions and for elucidating true TFAs. Similarly, others (114) explored a linear superposition of expression profiles and TFA combined appropriately using binding affinities in lieu of stoichiometric coefficients, and a Bayesian error analysis of an, effectively, linear method was presented in Reference 114. The main goal of this reverse-engineering is to identify the activation program of transcription modules under particular conditions (115) so as to hypothesize how activation/deactivation of expression can be induced/suppressed (116). A fundamental difference among the methods is whether the weights of the approximation should be estimated through regression (109–113) or associated with binding affinities (114). To gain more mechanistic insight, recent approaches aim at combining time-course expression data and semimechanistic models of gene expression in an effort to evaluate the kinetics of gene expression. In a report by Yugi et al. (117), a microarray data-based kinetic method (MASK) was proposed that combined expression data with RNA synthesis kinetic models to evaluate kinetics parameters of the expression activation and repression processes, whereas in a report by Thomas et al. (118), the so-called S-system framework was explored (119).

Truly fascinating, however, are the opportunities offered by combining time-course expression data with mechanistic-models of expression in the form of pharmacokinetic/pharmacodynamic expressions. Using gene arrays in the time-series paradigm can provide the scope of data necessary for analyzing dynamic complex biological phenomena. The time series can capture the dynamic nature of processes such as disease progression or drug responses, whereas the gene arrays provide a method of high-throughput data collection necessary to address the complexity. For years, complex pathologies such as diabetes, hypertension, and obesity have, for the most part, been addressed one or two genes at a time. Although such pathologies may be instigated by a single gene defect induced through gene knockout, in general, this is not the case. For example, a major animal model for obesity and related pathologies is the ZDF rat (120). This rat contains a single gene defect, the leptin receptor. In reality, this is not the human condition. Obesity and metabolic syndrome results from a complex interplay of many genes in multiple tissues (121). However, taking advantage of the opportunity to analyze such dynamic complex biological phenomena requires quantitative approaches that are able to accommodate both the dynamics and the scope of data. Indirect effect mathematical modeling provides an approach to addressing this problem. Although developed for pharmacokinetics and pharmaco-dynamics, the basic

approach can be applied to any dynamic biological system (122). The basic premise of such modeling is that a measured response (R) to an input perturbation may be produced by indirect mechanisms; for example, factors controlling the input or production of the response variable (k_{in}) may be either inhibited or stimulated, or the determinants of loss of the response variable (k_{out}) may be inhibited or stimulated. The rate of change of the response over time with no input perturbation present can be described by

$$\frac{dR}{dt} = k_{in} - k_{out} \cdot R \quad R(0) = R_0,$$

where k_{in} represents the zero-order constant for production of the response and k_{out} defines the first-order rate constant for loss of the response. It is assumed that k_{in} and k_{out} fully account for production and loss of the response. The response variable R may be a directly measured entity or an observed response, which is immediately proportional to the concentration of R. The basic assumption that both production and loss can be stimulated or inhibited leads to the four basic equations shown in **Figure 2**. The initial input perturbation is used as the driving force for the primary response or set of responses. However, the primary response(s) can then be employed as the driving force for a set of secondary responses. Using this approach, dynamic models for ever-more complicated converging and diverging sequences of molecular events in time can be constructed. For example, suppose a drug enhances the expression of two different transcription factors. These represent primary responses. If these transcription factors change the expression of other genes, then these become secondary responses. Changes brought by these genes become tertiary responses. In this way, the use of the four basic models can be used to construct experimentally testable models for quite complex response cascades. However, clustering of gene array time series data not only provides the foundation of such dynamic models but also determines their validity.

Acknowledgments

The authors would like to acknowledge insightful comments, suggestions, and guidance from Prof. W.J. Jusko and Prof. D.C. DuBois. I.P.A. and E.Y. acknowledge support from the National Science Foundation under an NSF-BES 0519563 Metabolic Engineering Grant and the Environmental Protection Agency under grant EPAGAD R 832721-101. R.R.A. acknowledges support by grants GM 24211 and GM 67650 from the National Institute of General Medical Sciences, NIH, Bethesda, MD, and by a grant from NASA.

LITERATURE CITED

1. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Molecular Biology of the Cell. Garland Sci; New York: 2002.
2. Kafatos FC. A revolutionary landscape: the restructuring of biology and its convergence with medicine. J. Mol. Biol. 2002; 319(4):861–67. [PubMed: 12079311]
3. Bowtell DD. Options available—from start to finish—for obtaining expression data by microarray. Nat. Genet. 1999; 21(Suppl. 1):25–32. [PubMed: 9915497]
4. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat. Genet. 1999; 21(Suppl. 1):33–37. [PubMed: 9915498]
5. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. Making and reading microarrays. Nat. Genet. 1999; 21(Suppl. 1):15–19. [PubMed: 9915495]

6. Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, et al. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 2003; 31(19):5676–84. [PubMed: 14500831]
7. Miklos GL, Maleszka R. Microarray reality checks in the context of a complex disease. Nat. Biotechnol. 2004; 22(5):615–21. [PubMed: 15122300]
8. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. Nat. Methods. 2005; 2(5):337–44. [PubMed: 15846360]
9. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. Multiple-laboratory comparison of microarray platforms. Nat. Methods. 2005; 2(5):345–50. [PubMed: 15846361]
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270(5235):467–70. [PubMed: 7569999]
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531–37. [PubMed: 10521349]
12. Jayaraman A, Yarmush ML, Roth CM. Evaluation of an in vitro model of hepatic inflammatory response by gene expression profiling. Tissue Eng. 2005; 11(1–2):50–63. [PubMed: 15738661]
13. Huang S, Eichler G, Bar-Yam Y, Ingber DE. Cell fates for high-dimensional attractor states of a complex regulatory network. Phys. Rev. Lett. 2005; 94:128701. [PubMed: 15903968]
14. Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, et al. Application of genome-wide expression analysis to human health and disease. Proc. Natl. Acad. Sci. USA. 2005; 102(13):4801–6. [PubMed: 15781863]
15. Eichler GS, Huang S, Ingber DE. Gene expression dynamics inspector (GEDI): for integrative analysis of expression profiles. Bioinformatics. 2003; 19(17):2321–22. [PubMed: 14630665]
16. Deleted in proof
17. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 2001; 7(6):673–79. [PubMed: 11385503]
18. Greer BT, Khan J. Diagnostic classification of cancer using DNA microarrays and artificial intelligence. Ann. NY Acad. Sci. 2004; 1020:49–66. [PubMed: 15208183]
19. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl. Acad. Sci. USA. 2002; 99(9):6163–68. [PubMed: 11983907]
20. Bar-Joseph Z. Analyzing time series gene expression data. Bioinformatics. 2004; 20(16):2493–503. [PubMed: 15130923]
21. Straume M. DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. Methods Enzymol. 2004; 383:149–66. [PubMed: 15063650]
22. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell. 2001; 106(6):697–708. [PubMed: 11572776]
23. Altenhein B, Becker A, Busold C, Beckmann B, Hoheisel JD, Technau GM. Expression profiling of glial genes during *Drosophila* embryogenesis. Dev. Biol. 2006; 296(2):545–60. [PubMed: 16762338]
24. Ko MS. Expression profiling of the mouse early embryo: reflections and perspectives. Dev. Dyn. 2006; 235(9):2437–48. [PubMed: 16739220]
25. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. A network-based analysis of systemic inflammation in humans. Nature. 2005; 437(7061):1032–37. [PubMed: 16136080]
26. Heller MJ. DNA microarray technology: devices, systems, and applications. Annu. Rev. Biomed. Eng. 2002; 4:129–53. [PubMed: 12117754]
27. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003; 4(9):117. [PubMed: 12952525]
28. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol. 1999; 19(3):1720–30. [PubMed: 10022859]

29. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between coexpression, coregulation and gene function. BMC Bioinform. 2004; 5:18.
30. Park PJ, Butte AJ, Kohane IS. Comparing expression profiles of genes with similar promoter regions. Bioinformatics. 2002; 18(12):1576–84. [PubMed: 12490441]
31. Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh E. Indexing multidimensional time-series. VLDB J. 2006; 15(1):1–20.
32. Jiang DX, Tang C, Zhang AD. Cluster analysis for gene expression data: a survey. IEEE Trans. Knowl. Data Eng. 2004; 16(11):1370–86.
33. Schliep A, Costa IG, Steinhoff C, Schonhuth A. Analyzing gene expression time-courses. IEEE/ACM Trans. Comp. Biol. Bioinform. 2005; 3(2):179–93.
34. Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans. Neural Netw. 2005; 16(3):645–78. [PubMed: 15940994]
35. D'Haeseleer P. How does gene expression clustering work? Nat. Biotechnol. 2005; 23(12):1499–501. [PubMed: 16333293]
36. Camastra F, Verri A. A novel kernel method for clustering. IEEE Trans. Pattern Anal. Mach. Intell. 2005; 27(5):801–5. [PubMed: 15875800]
37. Smola AJ, Scholkopf B. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. Algorithmica. 1998; 22(1–2):211–31.
38. Scholkopf B, Smola A, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 1998; 10(5):1299–19.
39. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA. 1998; 95(25):14863–68. [PubMed: 9843981]
40. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell. 2000; 11(12): 4241–57. [PubMed: 11102521]
41. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat. Genet. 1999; 22(3):281–85. [PubMed: 10391217]
42. Pan W, Lin J, Le CT. Model-based cluster analysis of microarray gene-expression data. Genome Biol. 2002; 3(2):RESEARCH0009. [PubMed: 11864371]
43. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. Bioinformatics. 2002; 18(2):275–86. [PubMed: 11847075]
44. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics. 2002; 18(3):413–22. [PubMed: 11934740]
45. Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. Proc. Natl. Acad. Sci. USA. 2002; 99(14):9121–26. [PubMed: 12082179]
46. Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. Dynamic modeling of gene expression data. Proc. Natl. Acad. Sci. USA. 2001; 98(4):1693–98. [PubMed: 11172013]
47. Schliep A, Steinhoff C, Schonhuth A. Robust inference of groups in gene expression time-courses using mixtures of HMMs. Bioinformatics. 2004; 20(Suppl. 1):I283–89. [PubMed: 15262810]
48. Schliep A, Schonhuth A, Steinhoff C. Using hidden Markov models to analyze gene expression time course data. Bioinformatics. 2003; 19(Suppl. 1):i255–63. [PubMed: 12855468]
49. Ji X, Li-Ling J, Sun Z. Mining gene expression data using a novel approach based on hidden Markov models. FEBS Lett. 2003; 542(1–3):125–31. [PubMed: 12729911]
50. Giurcaneanu CD, Tabus L, Astola J. Clustering time series gene expression data based on sum-of-exponentials fitting. Eurasip J. Appl. Signal Process. 2005; 2005(8):1159–73.
51. Vitanyi PMB, Li M. Minimum description length induction, Bayesianism, and Kolmogorov complexity. IEEE Trans. Informat. Theory. 2000; 46(2):446–64.
52. Akaike H. A new look at the statistical model identification. IEEE Trans. Automat. Control. 1974; AC-19:716–23.
53. Wu FX, Zhang WJ, Kusalik AJ. Dynamic model-based clustering for time-course gene expression data. J. Bioinform. Comput. Biol. 2005; 3(4):821–36. [PubMed: 16078363]
54. Wu FX, Zhang WJ, Kusalik AJ. Modeling gene expression from microarray expression data with state-space equations. Pac. Symp. Biocomput. 2004; 2004:581–92. [PubMed: 14992535]

55. Storey JD, Xiao WZ, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc. Natl. Acad. Sci. USA. 2005; 102(36):12837–42. [PubMed: 16141318]
56. Balasubramaniyan R, Hullermeier E, Weskamp N, Kamper J. Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics. 2005; 21(7):1069–77. [PubMed: 15513997]
57. Qian J, Dolled-Filhart M, Lin Y, Yu HY, Gerstein M. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J. Mol. Biol. 2001; 314(5):1053–66. [PubMed: 11743722]
58. Syeda-Mahmood, T. Clustering time-varying gene expression profiles using scale-space signals.. IEEE Comput. Soc. Bioinform. Conf. (CSB'03); San Jose, CA: IBM Almaden Res. Cent.; 2003. p. 48
59. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinform. 2006; 7(1):191.
60. Yang, E.; Berthiaume, F.; Yarmush, ML.; Androulakis, IP. An integrative systems biology approach for analyzing liver hypermetabolism.. Presented at 9th Int. Symp. Process Syst. Eng./ 16th Eur. Symp. Comput. Aided Process Eng.; Garmisch-Partenkirchen/Ger. 2006; Elsevier;
61. Lin, J.; Keogh, E.; Lonardi, S.; Chiu, B. Proc. 8th ACM SIGMOD Workshop Res. Issues Data Min. Knowl. Discov. San Diego, CA: 2003. A symbolic representation of time series, with implication for streaming algorithms..
62. Gower JC, Ross GJS. Minimum spanning trees and single linkage analysis. Appl. Stat. 1969; 18:54–64.
63. Xu Y, Olman V, Xu D. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. Bioinformatics. 2002; 18(4):536–45. [PubMed: 12016051]
64. Xu Y, Olman V, Xu D. Minimum spanning trees for gene expression data clustering. Genome Inform. 2001; 12:24–33. [PubMed: 11791221]
65. Ho TK, Basu M. Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. 2002; 24(3):289–300.
66. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biol. Bioinform. 2004; 1(1):24–45. [PubMed: 17048406]
67. Cheng Y, Church GM. Biclustering of expression data. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2000; 8:93–103. [PubMed: 10977070]
68. Heard NA, Holmes CC, Stephens DA, Hand DJ, Dimopoulos G. Bayesian coclustering of Anopheles gene expression time series: study of immune defense response to multiple experimental challenges. Proc. Natl. Acad. Sci. USA. 2005; 102(47):16939–44. [PubMed: 16287981]
69. Heard NA, Holmes CC, Stephens DA. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. J. Am. Stat. Assoc. 2006; 101(473):18–29.
70. Hirano S, Tsumoto S. Empirical evaluation of dissimilarity measures for time-series multiscale matching. Found. Intell. Syst. 2003; 2871:454–62.
71. Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003; 19(4):459–66. [PubMed: 12611800]
72. Thalamuthu A, Mukhopadhyay I, Zheng XJ, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. 2006; 22(19):2405–12. [PubMed: 16882653]
73. Steinbach, M.; Ertoz, L.; Kumar, V. Challenges of clustering high dimensional data.. In: Wille, LT., editor. New Vistas in Statistical Physics-Applications in Econophysics, Bioinformatics, and Pattern Recognition. Springer-Verlag; Berlin/New York: 2003.
74. Gershon D. Dealing with the data deluge. Nature. 2002; 416(6883):889–91.
75. Duin, RPW. Classifiers in almost empty spaces.. In: Sanfeliu, A.; Villanueva, JJ.; Vanrell, M.; Alquezar, R.; Kain, AK.; Kittler, J., editors. ICPR15 Proc. 15th Int. Conf. Pattern Recognit., Barcelona, Spain; Los Alamitos. 2000; IEEE Comput. Soc. Press; p. 1-7.

76. Ho TK. A data complexity analysis of comparative advantages of decision forest constructors. Pattern Anal. Appl. 2002; 5(2):102–12.
77. Thompson DM, King KR, Wieder KJ, Toner M, Yarmush ML, Jayaraman A. Dynamic gene expression profiling using a microfabricated living cell array. Anal. Chem. 2004; 76:4098–4103. [PubMed: 15253648]
78. Ganascia, JG.; Velcin, J. Concept. Struct. Work. Lect. Notes Comput. Sci. Vol. 3127. Springer-Verlag; Berlin: 2004. Clustering of conceptual graphs with sparse data.; p. 156-69.
79. Partsinevelos P, Agouris P, Stefanidis A. Reconstructing spatiotemporal trajectories from sparse data. Isprs J. Photogr. Remote Sensing. 2005; 60(1):3–16.
80. Velcin J, Canascia JG. Default clustering from sparse data sets. Symb. Quant. Approach. Reason. Uncertain. Lect. Notes Comput. Sci. 2005; 3571:968–79.
81. Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 1997; 19(2):153–58.
82. Dougherty ER. Small sample issues for microarray-based classification. Comp. Funct. Genomics. 2001; 2(1):28–34. [PubMed: 18628896]
83. Hwang D, Schmitt WA, Stephanopoulos G. Determination of minimum sample size and discriminatory expression patterns in microarray data. Bioinformatics. 2002; 18(9):1184–93. [PubMed: 12217910]
84. Yang, EH.; Androulakis, IP. Assessing the information content of short time series expression data.. Proc. 28th IEEE EMBS Annu. Int. Conf.; New York. 2006;
85. Pan W. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics. 2006; 22(7):795–801. [PubMed: 16434443]
86. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA. 2005; 102(43):15545–50. [PubMed: 16199517]
87. Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. Bioinformatics. 2002; 18(Suppl. 1):S145–54. [PubMed: 12169542]
88. Huang D, Pan W. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics. 2006; 22(10):1259–68. [PubMed: 16500932]
89. Qian Y, Zhang K, Lai W. Constraint-based graph clustering through node sequencing and partitioning. Adv. Knowl. Dis. Data Min. Lect. Notes Comput. Sci. 2004; 3056:41–51.
90. Tung, AKH.; Han, J.; Lakshmanan, LVS.; Ng, RT. Constraint-based clustering in large databases.. Proc. Int. Conf. Database Theory. Lect. Notes Comput. Sci.; Berlin. 2001; Springer-Verlag; p. 405-19.
91. Almon, RR.; DuBois, DC.; Jin, JY.; Yao, Z.; Hazra, A., et al. Development, analysis and use of pharmacogenomic time series for pharmacokinetic/pharmacodynamic modeling of multi-tissue polygenic responses to corticosteroids.. In: Barnes, LP., editor. New Research on Pharmacogenetics. Nova Sci; New York: 2006.
92. Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 2002; 12(10):1574–81. [PubMed: 12368250]
93. Bolshakova N, Azuaje F, Cunningham P. A knowledge-driven approach to cluster validity assessment. Bioinformatics. 2005; 21(10):2546–47. [PubMed: 15713738]
94. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics. 2001; 17(4):309–18. [PubMed: 11301299]
95. Handl J, Knowles J, Kell DB. Computational cluster validation in postgenomic data analysis. Bioinformatics. 2005; 21(15):3201–12. [PubMed: 15914541]
96. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 2004; 5(12):R98. [PubMed: 15575972]
97. Almon RR, DuBois DC, Jusko WJ. Corticosteroid-regulated genes in rat kidney: mining time series array data. Am. J. Physiol. Endocrinol. Metab. 2005; 289(5):E870–82. [PubMed: 15985454]

98. Almon RR, DuBois DC, Piel WH, Jusko WJ. The genomic response of skeletal muscle to methylprednisolone using microarrays: tailoring data mining to the structure of the pharmacogenomic time series. Pharmacogenomics. 2004; 5(5):525–52. [PubMed: 15212590]
99. Darnell JE Jr. Transcription factors as targets for cancer therapy. Nat. Rev. Cancer. 2002; 2(10): 740–49. [PubMed: 12360277]
100. Levy DE, Darnell JE Jr. Stats: Transcriptional control and biological impact. Nat. Rev. Mol. Cell. Biol. 2002; 3(9):651–62. [PubMed: 12209125]
101. Ruminy P, Gangmeux C, Claeyssens S, Scotte M, Daveau M, Salier MP. Gene transcription in hepatocytes during the acute phase of a systemic inflammation: from transcription factors to target genes. Inflamm. Res. 2001; 50(8):383–90. [PubMed: 11556518]
102. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. 2001; 409(6819):533–38. [PubMed: 11206552]
103. van Steensel B, Delrow J, Bussemaker HJ. Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. Proc. Natl. Acad. Sci. USA. 2003; 100(5): 2580–85. [PubMed: 12601174]
104. Alter O, Golub GH. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. Proc. Natl. Acad. Sci. USA. 2004; 101(47):16577–82. [PubMed: 15545604]
105. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ. Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol. 2004; 5(8):R56. [PubMed: 15287978]
106. Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinform. 2004; 5:31.
107. Boulesteix AL, Strimmer K. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. Theor. Biol. Med. Model. 2005; 2:23. [PubMed: 15978125]
108. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat. Genet. 2001; 27(2):167–71. [PubMed: 11175784]
109. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. USA. 2003; 100(26):15522–27. [PubMed: 14673099]
110. Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. Metab. Eng. 2005; 7(2):128–41. [PubMed: 15781421]
111. Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V, Liao JC. Network component analysis of *Escherichia coli* transcriptional regulation. Abstr. Pap. Am. Chem. Soc. 2004; 227:U216–17.
112. Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowhury V, Liao JC. Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. Proc. Natl. Acad. Sci. USA. 2004; 101(2):641–46. [PubMed: 14694202]
113. Kao KC, Tran LM, Liao JC. A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. J. Biol. Chem. 2005; 280(43):36079–87. [PubMed: 16141204]
114. Sun N, Carroll RJ, Zhao H. Bayesian error analysis model for reconstructing transcriptional regulatory networks. Proc. Natl. Acad. Sci. USA. 2006; 103(21):7988–93. [PubMed: 16702552]
115. Wang W, Cherry JM, Botstein D, Li H. A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA. 2002; 99(26):16893–98. [PubMed: 12482955]
116. Ng A, Bursteinas B, Gao Q, Mollison E, Zvelebil M. pSTIING: a ‘systems’ approach towards integrating signaling pathways, interaction and transcriptional regulatory networks in inflammation and cancer. Nucleic Acids Res. 2006; 34:D527–34. [PubMed: 16381926]
117. Yugi K, Nakayama Y, Kojima S, Kitayama T, Tomita M. A microarray data-based semikinetic method for predicting quantitative dynamics of genetic networks. BMC Bioinform. 2005; 6:299.

118. Thomas R, Mehrotra S, Papoutsakis ET, Hatzimanikatis V. A model-based optimization framework for the inference on gene regulatory networks from DNA array data. Bioinformatics. 2004; 20(17):3221–35. [PubMed: 15247105]

119. Savageau MA. A theory of alternative designs for biochemical control systems. Biomed. Biochim. Acta. 1985; 44(6):875–80. [PubMed: 4038287]

120. Janssen SW, Martens GJ, Sweep CG, Ross HA, Hermus AR. In Zucker diabetic fatty rats plasma leptin levels are correlated with plasma insulin levels rather than with body weight. Horm. Metab. Res. 1999; 31(11):610–15. [PubMed: 10598829]

121. Dandona P, Aljada A, Chaudhuri A, Mohanty P, Garg R. Metabolic syndrome: a comprehensive perspective based on interactions between obesity, diabetes, and inflammation. Circulation. 2005; 111(11):1448–54. [PubMed: 15781756]

122. Dayneka NL, Garg V, Jusko WJ. Comparison of four basic models of indirect pharmacodynamic responses. J. Pharmacokinet Biopharm. 1993; 21(4):457–78. [PubMed: 8133465]

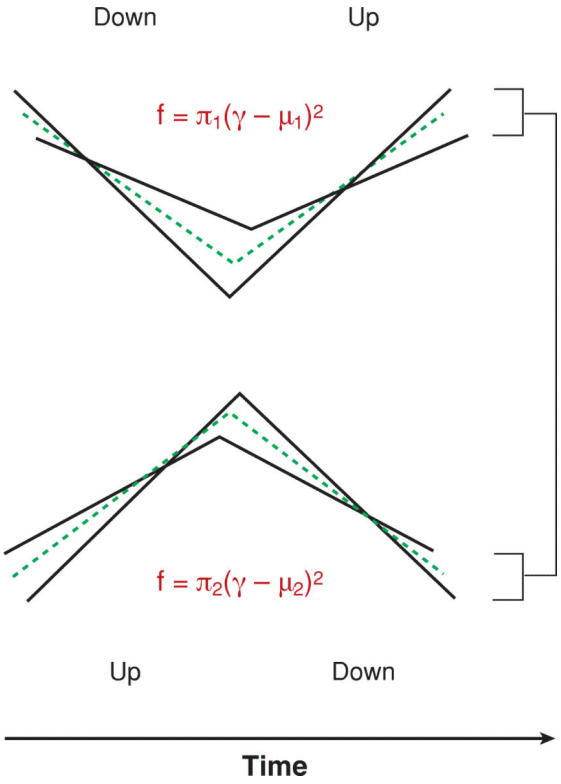

Figure 1.

Notional comparison of clustering methods. Given the four hypothetical trajectories, a distance-based method will compare the similarities pointwise, potentially creating an appropriate dendogram quantifying such distances. A model-based approach will attempt to quantify a functional description of the data in the form of a generalized model “f,” whereas a feature-based method will attempt to identify critical features, such as a sequence of events or trends, shared by various elements.

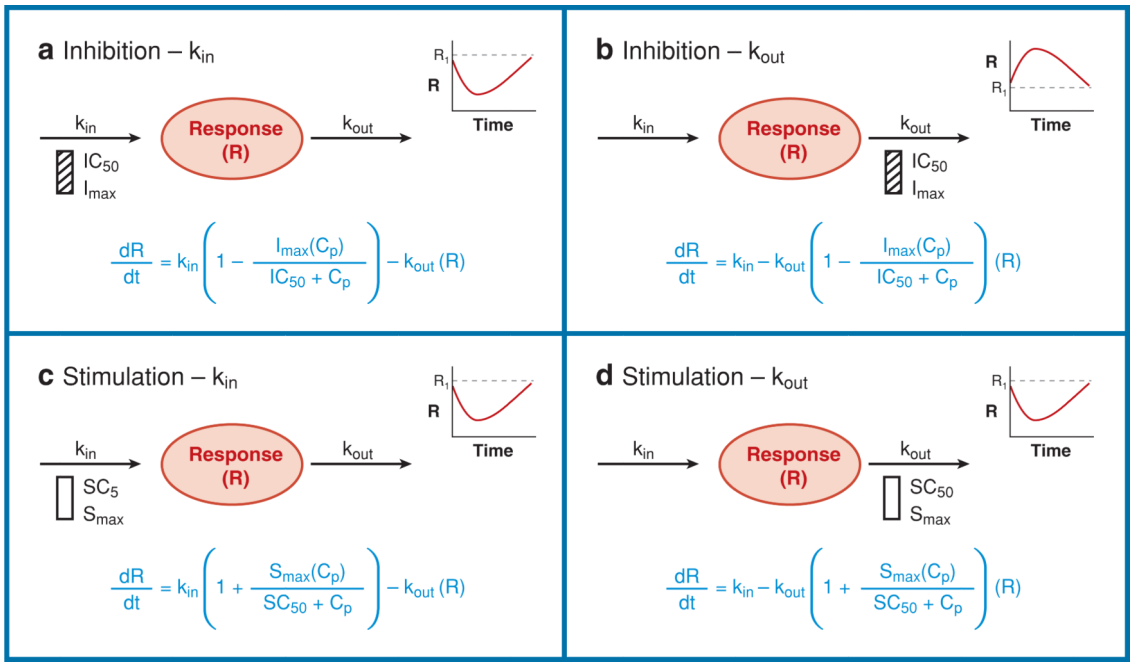

Figure 2.

Four basic mechanism-based indirect effect models for response dynamics indicating production and consumption inhibition, and production and consumption stimulation.

Table 1

Similarity metrics for PwDbM

L1-norm	$d_{ij} = \sum_{t=1}^{N_t} \mid E(i, t) - E(j, t) \mid$
2-norm	$d_{ij} = \sqrt{\sum_{t=1}^{N_t} (E(i, t) - E(j, t))^2}$
L∞-norm	$d_{ij} = \max_t \mid E(i, t) - E(j, t) \mid$
Mahalanoblis	$d_{ij} = (E_i - E_j)^T \Sigma^{-1} (E_i - E_j)$
Correlation metric	$d_{ij} = 1 - r_{ij}$ $r_{ij} = \frac{\sum_{t=1}^{N_t} \left(E(i, t) - \bar{E}(i)\right)\left(E(j, t) - \bar{E}(j)\right)}{\sqrt{\sum_{t=1}^{N_t} \left(E(i, t) - \bar{E}(i)\right)^2 \sum_{t=1}^{N_t} \left(E(j, t) - \bar{E}(j)\right)^2}}$