Post-Genomic Biology: Gene Expression Profiles, Cluster Analysis and Beyond

Sui Huang

Graft 2001; 4; 326

The online version of this article can be found at:
http://gft.sagepub.com

Published by:
SAGE Publications
http://www.sagepub.com

Additional services and information for Graft can be found at:
Email Alerts: http://gft.sagepub.com/cgi/alerts
Subscriptions: http://gft.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Post-Genomic Biology: Gene Expression Profiles, Cluster Analysis and Beyond

Sui Huang

With the completion of a first draft of the entire human genome sequence a new question has emerged: What does the “book of genes” tell us? Answering this question is the new challenge for the “post-genomic” era of biomedical sciences. A central technology that is helping us in this endeavor is the DNA microarray for the genome-wide analysis of gene expression. However, so far we have resorted to collecting, comparing and clustering genes and expression patterns. Given the quick spreading of these descriptive methods, the aim of this article is to serve as a primer to microarrays and to cluster analysis of gene expression data. We then discuss the future development of gene expression analysis toward an integrative view of the genome in the spirit of “biocomplexity.” Such a holistic, yet formal and data-based approach will be necessary if we are to learn the grammar in order to understand the meaning of the genome’s words.

Then, he holds the parts in his hand. But, missing is the spirit’s band.

—Mephisto (in Faust I, J. W. Goethe, 1808)

Introduction

Over the past centuries, the exploration of the very inner workings of the human body has suffered from an imbalance between the multitude of observed phenomena at the macro-level and the almost complete lack of knowledge of the details at the micro-level. This discrepancy has led to a continuous stream of ad hoc theories, ranging from the assumption of abstract “active principles” to the countless, carefully constructed arrow-arrow diagrams of pathogenesis pathways: We had too much theory but no molecular data.

Now, in the era of modern molecular medicine, we suddenly find ourselves facing a reverse situation: We have too much molecular data but no theory. As we stand at the threshold of the much celebrated “post-genomic” era of biology, ready to harvest the fruits of the human genome project, the central question now is: What theories will be appropriate to make sense of the excess of data?

The current flood of data in biology is mainly caused by two seminal technological achievements in genomic sciences:

1. automated DNA sequencing methods which have led to an explosion of gene and protein sequence information, culminating in a first draft of the entire human genome sequence, and

2. DNA microarrays which allow the simultaneous ("massively-parallel") measurement of the expression of thousands of defined genes.

The latter technology embodies the first step from traditional genomics, concerned with sequencing, toward “functional genomics,” which reflects our increasing awareness that DNA sequence alone does not provide an answer to the question of what individual genes “are good for.”

In exploiting the huge potential of DNA microarrays, we face the daunting complexity of the
living organism and the lack of a formal understanding of the genome-phenome relationship, and we have therefore resorted to picking the “low-hanging fruits” first; searching for genes that are associated with certain conditions, assigning a function to a gene and developing descriptive statistical methods to organize the continuing flurry of data. However, with an entire encyclopedia of all the genes of the human genome at hand, the next challenge will be to learn, from careful gene expression profile analysis, the grammar that is used by nature to integrate the words of the genome into the language of the phenome. We need a comprehensive, formal theory of genomic regulation and its functional consequences that goes beyond the current ad hoc assumptions of upstream/downstream relationships.

In their quest for a formal approach to deepen our knowledge of complex living systems, an increasing number of life scientists have teamed up with physicists and engineers to embark on a novel interdisciplinary, integrative scientific endeavor, perhaps best referred to as “biocomplexity,” to look beyond the gene sequence and molecular data. Their goal is to understand those “grammatical” rules that will lead to a conceptual framework that could fill the gap between genome and phenome, and hence between DNA sequence and disease, that we have sought to bridge with descriptive statistical associations.

Rather than presenting a comprehensive survey (which can be found in refs. 4-6), in this article we will give a practical overview on the principles of the methods currently used for generating and interpreting gene expression profiles. The central aim of this article is to serve as an introduction for clinicians and non-bioinformaticians to the basic idea of cluster analysis of expression data. Despite the aforementioned simplistic view behind it, clustering is now a widely used method for organizing genome-scale gene expression data and is at the threshold of entering the realm of clinical application. Thereafter, we discuss the need to overcome the current descriptive, gene-centered view that is rooted in traditional molecular biology and the need to develop integrative and formal approaches to the genome as a whole in the spirit of biocomplexity.

DNA Microarrays: Philosophy and Principle

In conventional methods for measuring expression of a gene (i.e., its mRNA level) in a tissue, such as Northern blots, a labeled nucleic acid probe recognizing the gene of interest is applied onto the sample RNA (containing all mRNA species) which is immobilized on a membrane. Such approaches entail on the investigator a prior decision as to which specific mRNA to examine. Thus, one needs a hypothesis of which gene could be interesting (i.e., expressed differently) in a given disease situation. In the absence of a hypothesis, one would wish to detect and identify a large number of mRNAs simultaneously ("discovery"- rather than "hypothesis"- driven science). However, this is impracticable given that typically only one non-discriminatory label (e.g., radioactivity or fluorometry) is available, which prevents distinctive tagging of multiple probes.

Microarray technology reverses the conventional approach to achieve simultaneous identification of a large set of nucleic acids (Fig. 1) by taking advantage of the availability of large collections of gene sequences. The individual DNA probes of known sequences are now immobilized as spots arranged in a dense array on a planar surface (glass or nylon) such that the grid coordinates of a spot now serve to encode the sequence identity of the immobilized DNA probe. The unknown mRNAs from the tissue sample to be analyzed are copied into cDNA target molecules during which they are uniformly labeled (with radioactivity or fluorescence). These labeled cDNA targets are then applied onto the entire array to allow hybridization with the tethered DNA probe. The arrayed spots capture the corresponding labeled cDNAs present in the sample and become labeled. After scanning the array for label activity, an image analysis software determines the pixel intensity of the spots, which reflects target abundance, and their position, which identifies the DNA probe. The final output is an electronic list (spreadsheet) of all the genes represented on the array with the value describing the abundance of the respective cDNAs in the interrogated sample (Fig.1, bottom). This procedure generates a genome-wide gene expression profile, the “transcriptome” (or a subset of it).

Unlike the genome, the transcriptome is a dynamic entity: It changes with time and environment within the framework defined by the static genome. With the development of protein arrays, the same ideas of genome-wide expression analysis are now moving to the protein level, enabling access to the proteome, which more accurately than the transcriptome reflects the actual activation state of the genes.
Figure 1. General principle of gene expression profiling with DNA microarrays. RNA is isolated from the sample (tissue or cells) and copied into cDNA, during which radioactive or fluorescent nucleotides are incorporated. (For linear amplification, the cDNA can be subjected to in vitro transcription to cRNA). The labeled single-stranded nucleic acid ("target"), representing the totality of expressed genes of the sample, is then incubated onto the microarrays for hybridization to the spotted "probe" DNA. Each spot on the array at a defined grid address contains a known DNA fragment acting as the probe. The illustration shows the hybridization of the three selected cDNA species, 3, 5, and 6, in the sample with the corresponding spot that contains the complementary DNA strand (inset). The spots that have bound the corresponding cDNA species in the sample become labeled. After washing the unbound excess cDNAs, the array is scanned to detect the label (radioactivity or fluorescence) and analyzed by image analysis software to measure the intensity value of each spot which is then translated (after background correction) to a numerical value that represents the expression level. The result is typically exported from the analysis program as a spreadsheet.
Other non-array-based techniques for monitoring the transcriptome include the re-sequencing of randomly concatenated fragments of cDNAs of the sample mRNA (serial analysis of gene expression), or oligonucleotide-tagged microbeads-based cDNA identification (massively-parallel signature sequencing). These methods are more laborious but have some application-specific advantages.

**DNA Microarrays: The Technology**

Physically, the idea of DNA microarrays can be implemented in various ways which differ mainly in the method of assembling the probe DNA with the solid carrier surface, in the use of either chemically synthesized oligonucleotides or larger (PCR-generated) DNA fragments, and in the way the hybridization event is detected.

1. **DNA Chips (Affymetrix’s “GeneChip” ®)**
   - The immobilized probes are oligonucleotides synthesized in situ using a photolithography technique. This allows one to pack approximately 10^7 nucleic acid sequences at high density in a small area of slightly more than 1 cm^2. Due to the shortness of the oligonucleotides (20-25 bases), a set of typically 20 spots of oligonucleotides corresponding to different sequences within the same gene is necessary to specify an mRNA. On a single currently available chip, up to 12,000 genes are represented. Detection of hybridization relies on fluorescence-labeling of the target molecules and confocal fluorescence microscopy-based scanning of the chips.

2. **DNA Microarrays (sensu stricto)**
   - Prefabricated DNA fragments (up to 1000 bp long) obtained from PCR amplification of clones from characterized cDNA libraries are micro-printed in a grid pattern onto a glass slide of the dimension of a conventional microscope slide. Several thousand different cDNAs can be spotted on a single slide. Detection of hybridized targets is typically based on fluorescence as described above. By labeling two different samples (e.g., disease vs. normal) with different color fluorochromes (e.g., red vs. green) and applying them simultaneously to the same array for competitive hybridization, a ratio for the relative expression of each gene in the two conditions can be obtained.

3. **DNA Nylon Filter Arrays (e.g., GeneFilter® from ResGen)**
   - Instead of glass, a nylon filter is used as a substrate for spotting cDNAs. Thus, this microarray method corresponds in principle to conventional dot blots. However, a dot is shrunk to the size of less than a millimeter, allowing around 5000 cDNA species to be printed onto a 5 × 7 cm array. The target is typically radioactively labeled. Unlike other techniques, the filters can be “stripped” and reused for other samples several times.

A direct comparison has shown that all three methods have similar absolute sensitivity, however, with regard to other practical aspects, such as signal-to-noise ratio, robustness, and minimal detectable relative mRNA abundance, the debate is still open. Typically, a change of expression level by factor of 2 can be reliably detected.

**Interpretation of Gene Expression Data**

We now describe various approaches used to extract information from gene expression profiling experiments in the historical order of increasing sophistication and integration. Emphasis is put on explaining cluster analysis using a pedagogical toy example of a simple 16-gene array interrogating four different samples (Fig. 2).

**Differential Expression in Two Situations**

An immediate application of massively-parallel hybridization with instant identification of the hybridizing genes is actually just an acceleration of conventional approaches, such as subtractive hybridization or differential display PCR. Thus, following the original goal of these laborious methods, a major, straightforward application of gene expression profiling is the identification of novel or known genes that are differentially expressed in distinct situations by comparing the population of mRNAs of two samples. For instance, one might be interested in finding genes that are differentially expressed in accepted or rejected allografts in order to identify the molecular players involved in tolerance or rejection. A typical use of microarray-based...
gene expression profiling thus far has been to compare healthy tissues with tumor tissues. Genes related to disease conditions, if shown to be specifically associated with, or even causative for a given condition, could then serve as a diagnostic marker or a therapeutic target, respectively.

Two-sample comparisons are typically presented as two-dimensional scatter plots in the \(xy\)-plane, where every dot represents a gene whose expression level in the two samples is indicated by the position with respect to the two axes \(x\) and \(y\). Fig. 3 shows two scatter plots derived from the pedagogical 16-gene example array of Fig. 2.

**Gene Clustering**

The relative ease of obtaining entire gene expression profiles of samples paved the way to broad-scale characterization of expression behavior for all the known genes in multiple conditions. A gene then is not characterized just by asking whether it is induced either in the control or the disease sample. Instead, the behavior of a large number, \(n\), of individual genes (where \(n\) can be in the thousands) is determined by their expression in a number, \(m\), of samples that can represent different patients, or cell cultures from a wide range of different conditions, such as the response to drugs, different cell types, or sim-
Cluster analysis is a statistical classification method to group objects (genes, patients, and so on) into "clusters" based on similarity of the object properties.

Co-regulated genes are genes that are induced or suppressed in the same situations (drug treatments, disease, tissue, and so on).

The new paradigm is that one looks at a whole group of genes rather than attempts to pick up some potentially pathognomonic genes. For instance, in a gene expression profiling study of retinas from diabetic rats, it was found that many genes that were up-regulated in diabetic retinopathy collectively suggested the presence of an inflammatory component, since they belonged to the family of leucocyte adhesion proteins and cytokines (Joussen et al, manuscript submitted). A similar study could give valuable insights into graft rejection models, for example by helping to classify types of rejection based on a "molecular signature" rather than histology.

Cluster analysis of genes is a tool that helps to find such biologically meaningful groups of genes. This is exemplified in our toy example shown in Fig. 2, which has \( n = 16 \) genes and \( m = 4 \) samples, yielding four expression profiles (A-D), each represented by 16 dots (for genes 1 to 16). The expression levels (dot intensity) can take, for sake of simplicity, just three different intensity levels. Since typical arrays with thousands of genes would lead to huge tables of numbers, the results are now routinely visually represented as tables of color-coded graphs.

The results of such an analysis (Fig. 3) show how genes are grouped into clusters based on their expression levels in different samples. The clusters can be used to infer biological functions or pathways that might be involved in the disease process.

Figure 3. Scatter plot graphs of comparative expression profiling between two samples. Two comparisons are shown based on the results of the toy example of Fig. 2. In the left scatter plot, samples A and C, and in the right scatter plot, samples A and B are compared. Each dot represents a gene whose expression level in the two samples of the comparison is expressed by the position of the dot. The unit of the axis is the expression level for a gene in the sample indicated by the axis label. In this example, only three discrete expression levels exist, indicated by the three sections for each axis. For illustration, a few genes are labeled individually (compare with results in Fig. 2). On such scatter plots, genes that are expressed at similar levels in both samples lie close to the diagonal of the graph, while differentially expressed genes appear as "outliers," away from the diagonal (e.g., genes 6 and 7 in the right panel). Note that samples A and C are more similar to each other (most genes are on the diagonal) than samples A and B. This is also apparent from visual inspection of the scanning results in Fig. 2.
fields as shown in Fig. 4 (top) for our 16-gene/4-sample example, in which a colored field represents the level of expression of a gene (“heat map”). Clustering aims at reorganizing the list of genes such that similarly behaving genes are grouped together. There is usually no optimal way for such re-grouping. A variety of cluster analysis algorithms exist—all with their specific, subtle advantages and shortcomings—which deliver similar but non-identical results (for details see refs. 22, 23). The field is still too new to allow meta-analysis and evaluation of the various methods for gene expression data analysis. Two basic types of clustering can be distinguished, hierarchical and non-hierarchical.

Hierarchical clustering. In hierarchical clustering, similarly behaving genes are grouped together, and the groups that behave similarly are grouped to larger clusters, giving rise to several hierarchical levels of relatedness (similarity) between the genes. The central notion is thus the similarity of the behavior of genes. Therefore, the first step in clustering is to establish a proximity matrix, a table containing a value for the mutual similarity between every pair among the n genes. Such a value is obtained using one of several ways to express the similarity or dissimilarity (distance) between two genes with respect to their expression behavior in the m samples (such as Euclidean distance, etc.). Then, a hierarchical tree (dendrogram), as we know from phylogenetic analysis, is built by joining the “closest” pair of genes (based on the similarity value in the proximity matrix) to a cluster and later the closest clusters into successively larger clusters, giving rise to several hierarchical levels (phase transition), following a pattern determined by the similarity matrix. The temperature range in which a cluster remains unchanged reflects the stability of that cluster—a measure that lacks in most of the agglomerative clustering methods.

Non-hierarchical clustering. Non-hierarchical clustering overcomes some of the problems that the hierarchical clustering methods inherently suffer from, such as the suggestion of a hierarchical descent, the lack of robustness, the generation of accidental structures and the lack of external control for the outcome of the cluster structure. Non-hierarchical clustering does not require the computation of the similarity table, which can become very large and slow down computation with the extended data sets that are expected in the near future. Non-hierarchical clustering, like the k-means method, requires the pre-specification of the number of clusters.21-23 In the self-organizing maps method (SOM),23 in addition to the cluster number, the geometry (e.g., rectangular grid configuration of the clusters) must also be specified. This allows one to impose some structure on the relationship between the clusters. SOM has been successfully applied to clustering mRNA profiles from yeast cell cycle and hematopoietic differentiation to extract characteristic temporal patterns of gene expression.26

Sample clustering. We have so far discussed how expressed genes can be clustered based on their be-
Figure 4. Pedagogic example for hierarchical clustering. Upper panel: the hybridization results of our 4-sample, 16-gene expression profile are represented in a typical “heat map,” a color-coded gene table of gene expression values (three levels of expression, taken from Fig. 2) in which the 16 genes are listed as rows and the 4 samples as columns. White = low, light grey = intermediate and dark grey = high level of expression. (In experiments in which samples are compared to a control sample, the colors usually denote relative decrease vs. increase of expression). Lower panel: the data set of the upper panel is subjected to hierarchical cluster analysis for both dimensions, the genes and the samples. The reordered table is shown with the dendrograms for the genes on the left and the samples on the top. Here, clustering of the genes yields two main clusters (a and b) with several subclusters. The degree of relatedness between the individual genes is reflected in the length of the branches and the number of the nodes (branching points) that separates them. For instance, genes 1 and 5 are very similar to each other in their behavior across the four samples, in that they differ only in the expression level in sample C. Note that the ordering of the clustered elements is somewhat arbitrary since for the same tree several possible orders of the gene rows exist. As for the sample clustering, the samples A and C form one cluster, and samples B and D form another cluster. The tree branches indicate that A and C are more closely related to each other than B and D. The results of the clustering can be verified by visual inspection of Fig. 2.
Two-way clustering treatment, tissue) is an object for clustering. Clustering of samples (i.e., of the m samples representing patients or conditions) rather than of the n genes. With a sufficient number of different samples from different conditions (large m), entire expression profiles representing individual samples can now serve as the objects (rather than as the properties of genes) for similarity-based clustering. Typically, n is 5,000-10,000 genes (the current capacity of microarrays) while m is in the range of 10 to 100 samples. Sample clustering is of particular clinical interest when the samples represent tissues derived from patients with a certain clinical condition. For instance, for colon adenocarcinoma, breast cancer, leukemia, lymphoma and melanoma, sample (patient) clustering has been reported to have revealed the diagnostic (sub)groups based on the gene expression profiles. Sample clustering can even lead to the discovery of new, distinct diagnostic subtypes that are of clinical relevance, as has been shown for diffuse large B cell lymphoma, which represents a large, heterogeneous group with respect to prognosis but has resisted further subclassification based on conventional morphological criteria.

Sample clustering is usually performed in combination with gene clustering on the same data matrix, leading to the color-coded tables (two-way or two-dimensional clustering) schematically illustrated in Fig. 4, bottom. In such a representation, the color-coded fields are typically arranged vertically according to gene clustering and horizontally according to sample clustering. While conventional two-way clustering is performed independently for the two dimensions (genes and samples), recently an elegant improvement, the “coupled two-way clustering,” has been proposed in which only particularly relevant subsets of genes (or samples) are used for clustering whereby the selection of the subsets relies on the results of the clustering itself. This iterative algorithm removes noisy data and has led to robust clusters and the discovery of partitions not found in conventional clustering.

**Supervised classification.** All the above methods of gene or sample clustering are blind procedures inasmuch as the resulting clusters (of genes or of samples) may or may not make biological sense and require interpretation in conjunction with experimental or clinical data. These methods are appropriate when no a priori knowledge is available and can lead to the discovery of novel, natural disease classes as discussed above.

Often, however, reference classes for genes or samples exist: functional families of genes, such as stress-induced genes, growth-related genes and so on, or clinical groups, such as non-tumor versus tumor samples, drug treatment versus untreated control, or graft rejection versus acceptance. Classification of expression profile data can then be framed as a problem of statistical pattern recognition and matching: For instance, the gene expression profile of a patient sample is to be matched with a set of known profiles that are characteristic of a specific condition. The practical relevance is obvious: “Give me a gene expression profile (of a patient) and I will tell you his/her condition.” Computational learning algorithms for such kinds of pattern recognition tasks have been developed for many technical applications in the industry. In supervised classification, say of patient biopsy samples, one needs a gold standard: a training set of samples (e.g., 50 patients) with confirmed diagnosis (e.g., malignant vs. benign tumors) is used to train the pattern recognition algorithm to distinguish between the two diagnoses. The algorithm can then be applied to the expression profiles of samples with unknown diagnoses which are then classified into the existing groups. Gene expression profile-based classification has first been demonstrated for the distinction between acute lymphoblastic and myeloid leukemia.

A promising machine learning method is the support vector machine (SVM) which has been trained to discriminate tumor from non-tumor tissue in colon adenocarcinoma samples. A future use of such gene expression profile analysis in transplantation is technically well within reach but would require systematic studies, that is, the ac-
GENETIC NETWORK:
The schematic representation of all the regulatory interactions (induction/inhibition) between the genes of the genome.

GENE EXPRESSION SPACE:
Abstract multidimensional space in which each dimension represents a gene, and the position on the respective axis determines the level of expression. Thus, a single point in state space represents a distinct cellular state with respect to gene expression.

Beyond Clustering: Integrative, “Complex System” Approaches
Disease classification and diagnosis based on pattern recognition of entire gene expression profiles represent a significant paradigm shift, away from traditional gene-centered molecular biology toward an increasingly integrated view (Fig. 5). It is recognized that sometimes not individual genes but the “expression signature” of an entire set of genes defines a diagnosis. It might even well be that the defining characteristic of a disease at the molecular level is the membership of its expression profile in a cluster of highly similar expression patterns in “gene expression space” rather than the presence of one single, obligate (pathognomonic) individual molecular marker.

The methods discussed so far essentially use “brute-force” statistical pattern recognition to identify clusters and find correlations of genes or gene expression profiles with clinical conditions. As stated above, this correlative information, based on the idea of guilt by association, can indeed have immediate practical significance. However, the clustering approaches do not provide a deeper conceptual understanding of the regulatory principles inherent in the expression profiles.

A fundamental biological question is: Why do we have clusters of genes and of entire samples at all? The patterns seen in genome-wide profiles of gene expression must somehow reflect the interaction between the genes. As discussed earlier, co-regulated genes will be expressed or suppressed in the same situations. In contrast, a pair of genes, A and B, will not be expressed simultaneously at high levels in the same sample if gene A is an unconditional inhibitor of gene B. Because of such regulatory interactions, the expression levels of individual genes cannot change independently of each other. An expression profile is thus a snapshot of the highly constrained dynamics and rule-like behavior of the ensemble of genes imposed by the underlying genomic regulatory network. The dynamic constraints massively reduce the number of theoretically possible gene expression profiles down to a relatively small number of typical, robust and biologically meaningful profiles. They also establish the rules that force changes of cellular phenotype into relatively few coherent processes, such as differentiation, proliferation, apoptosis, migration, cellular activation, and so on.27 The gene interactions and ensuing rule-like dynamics of cell behavior therefore represent the context and a grammar that gives the words of the genome a meaning. Can we exploit these dynamic constraints of expression profiles to read more information out of the observed patterns?

Reverse Engineering the Wiring Diagram of Regulatory Pathways
An obvious aim in gaining a deeper insight into regulatory principles is to reverse engineer (infer) the wiring diagram (map of all gene-gene interactions) of the entire genetic network based on the information inherent in the temporal profiles of gene expression across the genome. Theoretical work using idealized model networks has demonstrated the principles (and limits) of algorithms for such network inference.4 Due to practical problems, such as cellular heterogeneity of the samples and intrinsic as well as experimental noise in gene expression, the application of these achievements in idealized networks to real gene expression data is facing some challenges.

Global Dynamics of Gene Expression and Cell Behavior
Mapping out the gene-gene connections of the genomic regulatory network does not directly address the fundamental question about the rules through which the encoded genomic information is translated into the observable phenotype. Having a detailed wiring diagram at hand certainly would stimulate exciting and useful in silico experiments and thus facilitate the progress of our understanding of the organism at a higher level of organization. However,
philosophers of science would contend that being able to re-enact the full complexity of a real organism as a 1:1 model in the computer is not equal to a conceptual understanding of that complex system. The very existence of clusters of expression profiles that represent phenotypes such as cell fate, cell type, response to external perturbations or disease conditions points to a robust, self-organizing behavior that emerges from the collective action of the genes. The use of highly simplified models of the genetic network has been instrumental in providing insights into the fundamental, generic principles of how a large system of interacting genes gives rise to its “global observable,” the cell behavior with the constraints.\textsuperscript{27,38-41} The theoretical treatment of large model networks of anonymous genes suggests that the global dynamics of gene-gene interactions, as defined by the wiring diagram of the network, can be represented as a smooth landscape (in the abstract, high-dimensional gene expression space) which contains multiple attractors that act like energy minima.\textsuperscript{27,41} In that attractor landscape, a given gene expression profile at a given time, representing a cell state, is just a single point. Biological processes driven by the change of the expression profile can then be pictured as a marble (whose position represents the cell state) that is forced to roll along specified paths, the valleys, into attractor basins and end up in one of the attractors. The latter corresponds to a stable cellular phenotype, such as proliferation, differentiation or functionally activated state.\textsuperscript{41} The structure of the attractor landscape which is defined by the genomic wiring diagram determines how the gene expression profile changes; thus, it embodies the dynamic constraints and rule-like behavior of physiological and pathological cell processes, including self-stabilization, directionality and determination. In this model, environmental influences affect the position and course of the cell state by hitting the marble, which then has to roll in accordance with the attractor landscape sculpted by the genome. Thus, the model unites genetic determination with environmental inputs within one formal framework.

Although these integrating concepts are still in their infancy, they will provide valuable guidance in exploiting the marvelous post-genomic technologies to acquire genome-wide data in a more systematic and purposeful way. This novel union of formal theory with high-throughput data acquisition...
tion in biomedical research will enable us to learn how the parts, the gene sequences that we have in our hand, integrate to a whole, the living organism that we observe. This in turn could pave the way toward a holistic, yet molecular understanding of disease processes.

References