

# Computational Biology and Drug Discovery: From Single-Target to Network Drugs

Alberto Ambesi-Impombato<sup>1,2</sup> and Diego di Bernardo<sup>\*,1</sup>

<sup>1</sup>Telethon Institute of Genetics and Medicine (TIGEM), Via P. Castellino 111, 80131, Naples, Italy

<sup>2</sup>Department of Neuroscience, University of Naples "Federico II", Via Pansini 5, 80131, Naples, Italy

**Abstract:** The drug discovery process is complex, time consuming and expensive, and includes preclinical and clinical phases. The pharmaceutical industry is moving from a symptomatic relief focus towards a more pathology-based approach where a better understanding of the pathophysiology should help deliver drugs whose targets are involved in the causative processes underlying the disease. Computational biology and bioinformatics have the potential not only to speed up the drug discovery process, thus reducing the costs, but also to change the way drugs are designed. In this review we focus on the different computational and bioinformatics approaches that have been proposed and applied to the different steps involved in the drug development process. The development of 'network-reconstruction' methods is now making it possible to infer a detailed map of the regulatory circuit among genes, proteins and metabolites. It is likely that the development of these technologies will radically change, in the next decades, the drug discovery process, as we know it today.

## 1. INTRODUCTION

The drug discovery process is complex, time consuming and very expensive. Typically, the time to develop a candidate drug is about 5 years, while the clinical phases leading, possibly, to the commercial availability of the drug are even longer (>7 years) for a total cost of more than 700 Million dollars [1]. The drug discovery process begins from the identification of an area of "unmet medical needs" and then proceeds by identifying "druggable" biological targets that could relieve the symptoms of the disease, or, as in the recent years, that are involved in the causative process of the disease. The pharmaceutical industry is moving from a symptomatic relief focus towards a more pathology-based approach where a better understanding of the pathophysiology should help deliver drugs whose targets are directly involved in the causative processes underlying the disease [2]. The drug discovery process is very similar across different pharmaceutical companies. It consists of preclinical and clinical phases. In the *target identification and validation* step, "druggable" biological targets are identified. In the *hit identification* step, library of compounds ranging from tens to hundreds of thousands of compounds are screened against the "druggable" targets to identify those compounds that "hit" the targets using high throughput screening (HTS). HTS methods based on experimental assays are reviewed extensively elsewhere [3]. The number of compounds selected after this step is in the order of hundreds. By analyzing the structure of the selected compounds and identifying common active substructures, novel compounds containing those substructures are synthesized to significantly lower the number of lead compounds. This step is called *lead identification*. Structural

bioinformatics and chemical informatics approaches to drug discovery are particularly useful in this step, however, widely used methods like structure-activity relationship (SAR) are outside the scope of this review. We refer the interested reader to Bredel *et al.* [4] and Fagan *et al.* [5].

The leads identified are further refined to comply with pharmacokinetic constraints such as absorption and bioavailability, and to increase their potency and efficacy, while decreasing side effects and toxicity. This step is called *lead optimization*. Knowledge of the mode of action (MOA), that is, the identification of the therapeutic molecular target of the drug, can simplify the task of optimizing the drug candidate. Understanding the MOA can help predicting the effect of drug interactions and allow structure-activity relationships (SAR) to guide medicinal chemistry efforts toward optimization [3]. However, for many drugs, the targets are unknown and difficult to find among the thousands of gene products in a typical genome.

Many new compounds fail when they are tested in humans due to lack of efficacy. Testing for efficacy early during the drug discovery process (*i.e.* before the clinical phases) is essential for reducing costs and time required. Therefore, the development of experimental and computational approaches to test for efficacy *in vitro* is critical.

After the preclinical phases, a candidate compound is then selected and the clinical phase of the process can begin. This consists of clinical phase I, phase II, and phase III and possibly the launch into the market. Many compounds fail in the clinical phases of the process thus leading to consistent waste of time and money. A good review of the evolution of the drug discovery process can be found in Ratti *et al.* [2].

Computational biology and bioinformatics have the potential not only of speeding up the drug discovery process thus reducing the costs, but also of changing the way drugs are designed. In this review we will focus on the different computational and bioinformatics approaches that have been

\*Address correspondence to this author at the Telethon Institute of Genetics and Medicine (TIGEM), Via P. Castellino 111, 80131, Naples, Italy; Tel: +39 081 6132 319; Fax: +39 081 6132 351; E-mail: dibernardo@tigem.it

**Table 1. Classification of the Reviewed Manuscripts According to the Computational Methods Used and their Application to the Drug Discovery Process**

Drug Discovery		Classifiers	Network/Pathway Reconstruction
Target identification & validation		Stoughton <i>et al.</i> 2005 (Review) [8]; Walker 2001 (Review) [47]; Hughes <i>et al.</i> 2000 [9]; Gasch 2000 [11]; Stegmeir 2004 [13]; Brown 2000 [12]	Gardner <i>et al.</i> 2003 [37]; Basso <i>et al.</i> 2005 [38]; Gardner <i>et al.</i> 2005 (Review) [35]; Apic 2005 (Review) [34]
Hit identification, Lead identification & optimization	Mode of action (MOA)	Perlman <i>et al.</i> 2004; Parsons <i>et al.</i> 2003 [48]; Parsons <i>et al.</i> 2004 [18]; Marton 1998 [17]; Giaever 2004 [20]; Giaever 1999 [19]; Lum 2005 [22]; Hughes <i>et al.</i> 2000 [9]; Betts 2003 [24]; Paull 1989 [14]; Weinstein 1997 [15]; Bao <i>et al.</i> 2002 [16]	di Bernardo <i>et al.</i> 2005 [42]; Imoto 2003 [39]; Haggarty 2003 [41]
	Efficacy & Toxicity	Bugrim <i>et al.</i> 2004 (Review) [49], Szakacs 2004 [27]; Staunton <i>et al.</i> 2001 [29]; Scherf 2000 [25], Gunther 2003 [30]; Gunther 2005 [31]; Hamadeh 2002 [50,51]; Dan <i>et al.</i> 2002 [26]	Not known

proposed and applied to the different steps involved in drug development as shown in Table 1. Our aim is to describe the different computational methods that have been used so far to tackle these problems by giving examples of applications. Since we cannot be comprehensive in our review, we tried to compensate for this by referring the interested readers to other reviews with a different focus that have been written on this subject. The organization of this paper is based on classifying drug discovery approaches into two major categories. Section 2 reviews Classifier-based algorithms which try to determine drug specific patterns as biomarkers of a compound activity, while section 3 assesses more complex methods that attempt to infer the network of gene-gene interactions that are perturbed by a drug. We further subdivided those sections in subsections, each focusing on specific steps of the drug discovery process.

## 2. CLASSIFIER-BASED ALGORITHMS

A classifier is an algorithm that uses a set of input or predictor variables  $x = (x_1, x_2, \dots, x_n)$  to predict one or more response variables  $y = (y_1, y_2, \dots, y_m)$  (Fig. 1). For example  $x$  can be a set of measurements of the expression of  $n$  genes in response to a drug treatment in a tumor cell type and  $y$  can represent the efficacy of the drug for that tumor cell type. Classifiers can be further subdivided in supervised-learning methods and unsupervised-learning methods. In supervised-learning a training set of ‘solved cases’ is used to train a model to recognize what will be the response  $y$  given the input variables  $x$ . Supervised-learning methods may be thought of as a “learning with a teacher model” in which a student gives an answer  $\hat{y}$  to each question  $x$  in the training set, and the teacher provides the correct answer  $y$ . After the training, the student should be able to give the correct answer to a new question that was not in the training set. If  $y$  and  $\hat{y}$  are coded as numerical values, we can define a loss function  $L(y, \hat{y})$ , for example,  $L(y, \hat{y}) = (y - \hat{y}(\theta))^2$ , where  $\theta$  are the parameters of the model to be learned. By minimizing this function over  $\theta$  on the training set, one finds the values of the model parameters  $\theta$ . For example, Linear Discriminant Analysis (LDA) is a supervised learning where  $\hat{y} = \theta x$ .

In unsupervised-learning, or “learning without a teacher”, one has a set of  $n$  observations  $(x_1, x_2, \dots, x_n)$  without the correct response variables. Cluster analysis is an example of unsupervised-learning method whose goal is to group a

collection of objects into subsets or “clusters”, such that the objects within each cluster are more closely related to one another than those assigned to different clusters. In addition the goal can also be to arrange the clusters in a natural hierarchy. A commonly used hierarchical clustering is the one described by Eisen [6]. Unsupervised methods have the advantage that they are ‘data driven’ and do not rely on *a priori* knowledge. A comprehensive and detailed description of these methods can be found in the excellent book by Hastie *et al.* [7].

### 2.1. Target Identification and Validation

Whole-genome gene expression data, proteomic data or metabolomic data, also named “molecular profiling” in a recent review [8], can be used to build classifier algorithms able to help in the process of identifying ‘druggable’ gene/protein/metabolites targets.

An example of an unsupervised-learning method can be found in Hughes *et al.* [9]. These authors constructed a reference database of whole-genome expression profiles referred to as a gene expression “compendium” generated by 300 diverse mutations and chemical treatments in *Saccharomyces cerevisiae*. A 2D hierarchical clustering [6,10] was used to cluster genes and experiments using as the similarity measure the correlation coefficient. Genes and experiments were reordered according to the resulting clustering similarity trees. By examining the clusters the authors were able to find an unknown ORFs that clustered among genes involved in the ergosterol biosynthesis and experiments that were perturbing this pathway, thus deducing these ORFs to belong to this pathway. They then experimentally confirmed that 8 of these ORFs were indeed required for sterol metabolism. Since sterol metabolism is a ‘druggable’ pathway in yeast for antimycotic drugs, this work shows how novel targets can be identified *via* bioinformatics approaches. A similar method has been applied by Gasch *et al.* [11] that performed a hierarchical clustering of 142 whole-genome arrays in *S. cerevisiae* in response to environmental changes and were able to clarify the regulation mechanisms in which three transcription factors were involved.

An example of supervised learning for understanding the function of gene from gene expression data is given in Brown *et al.* 2000 [12], in which Support Vector Machines

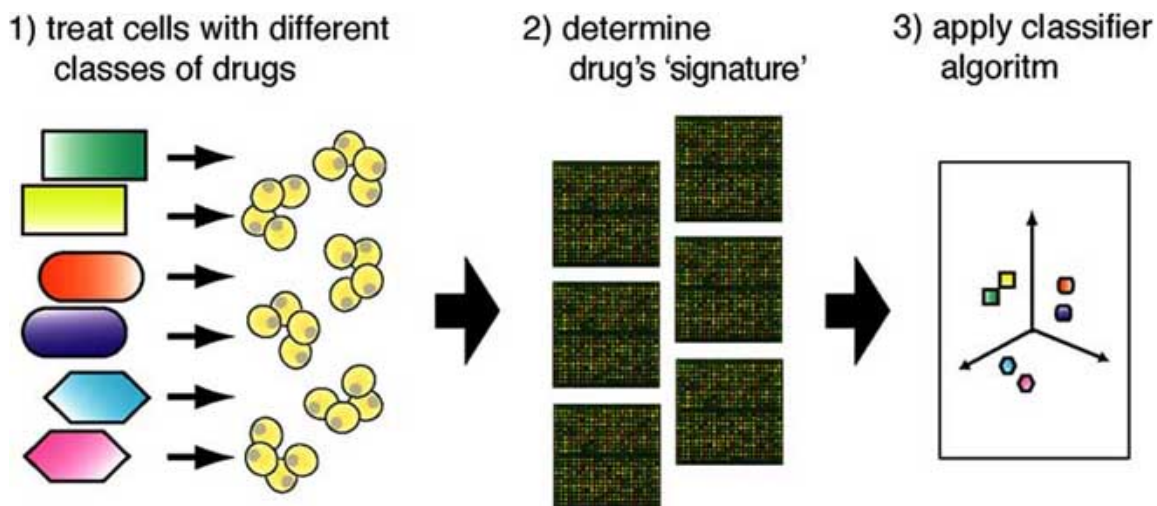


Fig. (1). Schematic diagram of the classifier-based algorithms.

(SVMs) [7] are used. When applied to gene expression data, an SVM begins with a set of genes that have a common function: for example, genes coding for ribosomal proteins or genes coding for components of the proteasome. In addition, a separate set of genes that are known not to be members of the functional class is specified. These two sets of genes are combined to form a set of training examples in which the genes are labeled positively if they are in the functional class and are labeled negatively if they are known not to be in the functional class. By analyzing expression data from 2,467 genes from the budding yeast *S. cerevisiae* measured in 79 different DNA microarray hybridization experiments the authors were able to correctly assign genes to five functional classes from the Munich Information Center for Protein Sequences Yeast Genome Database (MYGD). The method is compared with hierarchical clustering and shown to be marginally better, but this could have been expected since supervised learning methods have access to additional information as provided by the training set.

An original high throughput drug screening strategy based on unsupervised-learning is used by Segmaier *et al.* [13], which, unlike most commonly used methods, does not simply screen for compounds that interact with specific molecular targets. The authors preliminarily define a gene expression signature for the target post-treatment phenotype, or 'cellular state' of interest. Specifically, in this study the target cellular state was differentiated neutrophils and monocytes from control individuals *vs.* pretreatment bone marrow samples derived from Acute Myelogenous Leukemia (AML) patients. A "handful" of marker genes were selected, unfortunately not in a generalized manner but rather arbitrarily, from the differentiation-correlated genes. Those markers were then used to develop a detection assay called Gene Expression-based High-Throughput Screening (GE-HTS) based on multiplexed RT-PCR and Single Base Extension (SBE) reaction followed by MALDI-TOF mass spectrometry. Eight target compounds identified by GE-HTS in this study were validated in several ways including

morphological observations and functional measures. Interestingly, the broader cellular genetic program of differentiation beyond the selected handful of marker genes was also investigated, again through a correlation-based statistical test. The authors analyzed triplicate microarray expression data from HL-60 cell lines treated with eight different compounds. Six out of the eight expression profiles were found statistically significantly similar to the gene expression differences characterizing the original AML-*vs.*-controls primary cells, as determined by the Mantel test [13]. This test is an unbiased, global measure of similarity, and indicates that the six compounds induced a nonrandom pattern of gene expression consistent with differentiation. The advantage of GE-HTS is that the development of the assay does not require any specialized assays such as traditional methods based on antibodies or reporter constructs or cellular phenotypes, and, once the gene expression signature pattern is defined, the procedure is rather straightforward.

Although they may result in outstanding accuracy performance, correlation-based methods do not easily provide insight into the mechanisms of action common to the therapeutic category, but rather capture silent features of drug efficacy by their correlation to biomarker signatures based on gene expression patterns.

## 2.2. Hit Identification, Lead Identification and Optimization: Mode of Action (MOA)

One of the first bioinformatics approaches to determine mode of action of a compound was based on a simple supervised-learning approach [14]. In 1985 the National Cancer Institute (USA) established a primary screen in which compounds were tested *in vitro* for their ability to inhibit growth of 60 different human cancer cell lines [15]. To each compound tested it is possible to associate a value quantifying the differential growth inhibition (GI) for each cell line (treated *vs.* untreated). The algorithm developed by Allen and coworkers, named COMPARE, measures the similarity of the GI "signature" of a novel compound against

a database of “signatures” of compounds with known MOA. The similarity is obtained simply by computing the average differences between the signatures of the test compound and each of the signatures in the database. Ranking according to this measure of similarity, one can infer the MOA of the novel compound as the one of the most similar compound in the database. An extension of this approach based on hierarchical clustering and integration of different data set from the NCI 60 cell lines has been proposed by Weinstein in 1997 [15]. A more sophisticated approach using SVM to classify drugs into 5 mechanistic classes using drug activity profiles and the gene expression profiles of each of the untreated NCI 60 cell lines, has been proposed by Bao *et al.* [16].

Unsupervised approaches have been applied extensively in this area. Marton *et al.* [17] were pioneers of the “signature approach” based on gene expression profile following drug treatment. In this approach the drug signature is compared to a mutant strain signature using a correlation coefficient as a measure of similarity,  $p = \frac{\sum x_k y_k}{\sqrt{\sum x_k^2 \sum y_k^2}}$ . They

also proposed a further ‘decoder’ step where the mutant strains whose expression profiles were most similar to the drug-treated cells are treated with the drug, generating an expression signature in the mutant strain. If the mutated gene encodes a protein involved in the pathway affected by the drug, then the signature in mutant cell should be different or, ideally, absent. Marton *et al.* did a proof of principle study on FK506 and the calcineurin signaling pathway as a model system.

The previously described work by Hughes *et al.* [9], is another good example of how hierarchical clustering and correlation can be used for understanding the MOA of a drug. The authors used the gene expression ‘compendium’ to identify the target of the commonly used topical anesthetic dyclonine. In order to find the target of the compound, the authors treated the yeast cells with the compound and compared the gene expression profile to the most similar expression profiles in the compendium using the correlation coefficient as the similarity measure. The *erg2*  $\Delta$  strain (knock-out of the *erg2* gene) was most similar to the dyclonine treatment thus suggesting, correctly as verified experimentally, that this gene is the molecular target of the drug. Since this gene is conserved in human but codes for the sigma receptor, a neurosteroid-interacting protein, the MOA of the drug in human has also been explained.

Hierarchical clustering methods have been applied not only to gene expression data, but also to chemical-genetic and genetic interaction data. Parsons *et al.* [18] screened ~4700 yeast deletion mutants for hypersensitivity to 12 diverse inhibitory compounds. Hypersensitivity was measured from digital images of plates by quantifying colony area growing in drug-medium versus no-drug control medium. Hypersensitive strains for a given drug were coded as 1, and with a 0 otherwise. These data (a vector of ~4700 0s and 1s for each drug) were used for 2D hierarchical clustering. Both genes and compounds are clustered together upon the similarity of their chemical-genetic interactions. By analyzing the clusters they were able to detect genes whose

deletion was associated with sensitivity to multiple compounds, thus enabling them to identify a multidrug-resistant gene set. To identify the mode of action of a compound, they performed synthetic lethal screens between ERG11 mutants and the ~4700 deletion strains. The overlap between the genes that were synthetic lethal with ERG11 mutants, with the genes whose deletions were lethal after treatment with flucanazole, was used to infer the MOA of this drug.

Related to these methods are drug-induced haploinsufficiency screens first proposed by Giaever *et al.* [19]. Drug-induced haploinsufficiency occurs when lowering the dosage of a single gene from two copies to one copy in diploid cells results in a heterozygote that displays increased sensitivity to the drug as compared to the wild-type strain. These screens make use of a fitness defect score [20] that is computed using different methods [21,22].

Hierarchical clustering has been applied also to data derived from automated microscopy in order to identify drug MOA. Perlman *et al.* [23] chose 200 compounds, 90 of which were drugs with known MOA. They cultured HeLa (human cancer) cells in 384-well plates to near confluence, and treated them with 13 threefold dilutions of each drug for 20 hour, covering a final concentration range from micromolar to picomolar. They chose 11 distinct fluorescent probes covering a range of biological processes. Using automated fluorescence microscopy they measured for each cell, region and probe, a set of descriptors including size, shape, intensity, as well as ratios of intensities between regions for a total of 93 descriptors. For each descriptor they developed a titration-invariant similarity score (TISS) to allow comparison between dose-response profiles independent of starting dose. TISS scores for 61 compounds were computed and used for hierarchical clustering; the data matrix used for clustering consisted of 61 compounds by 93 TISS scores. Once again they found that drug with similar mechanism of action clustered together, thus allowing inference of drug MOA for drugs with unknown molecular targets.

Signature Expression profiles were used by Betts [24] to determine the differential mode of action of three active drugs against *Mycobacterium tuberculosis*, and as a means of identifying novel and efficacy-optimized active drugs. In this study the authors show that although global response profiles of isoniazide and thiolactomycine are more closely related to each other than to that of triclosan, there are differences that distinguish the mode of action of these two drugs. A mathematical model is proposed to discriminate between the three compounds and also the vehicle control treatment. The main sources of variance of the data were obtained by Principal Component Analysis (PCA). The principal components are a linear combination of all the gene intensities. Partial least squares discriminant analysis was performed on a subset of data selecting the dose and the time point that maximized separation of experimental groups. The 500-top ranking genes thus identified, were further processed by stepwise linear discriminant analysis in order to generate a mathematical model for the probability  $P_i(x)$  of a gene expression signature  $x$  belonging to classification group  $i$  based on the following discriminant function:

$$P_i(x) = \frac{e^{D_i^2(x)}}{\sum_{j=1}^n e^{D_j^2(x)}} \quad i = 1, 2, \dots, n \quad (1)$$

where  $D_i^2(x)$  is the discriminant score of the signature  $x$  for group  $i$ .

Methods that rely on a dataset for the construction of a classifier model, without implementing more robust statistical analyses, such as running a series of training and testing data in a 'leave-one-out' manner, although accurately performing on the training dataset may lead to the construction of a model that 'overfits' the data, and thus may not perform well on new data obtained using different treatments.

### 2.3. Hit Identification, Lead Identification and Optimization: Efficacy and Toxicity

A large part of the efforts based on computational and bioinformatics approaches have been directed to predict sensitivity of cancer cell lines to different compounds. Sherf *et al.* [25] aimed at relating sensitivity to therapy with gene expression using an unsupervised approach. They used the database of drug activity profiles (Growth Inhibition after 48h of drug treatment) of more than 70,000 compounds on NCI 60 cell lines, together with gene expression profiles of 9,703 genes measured using cDNA microarrays for each of the 60 untreated cell lines. They then performed a hierarchical clustering of 118 compounds with known mechanism of action. In order to integrate drug activity profile with gene expression data, they chose Pearson correlation coefficient as a measure of similarity. This coefficient was calculated for each combination of a gene (expression profile across 60 cell lines) and a drug (GI activity profile across 60 cell lines). This yielded 1376 correlation coefficients for each of the 118 drugs. Using this technique they were able to associate sensitivity of leukemic lines to L-asparagine to the amount of asparagine synthetase. A similar technique has been proposed by Dan *et al.* [26]. They were able to identify gene markers for chemosensitivity for 55 anticancer drugs using gene expression data across 39 cell lines and drug activity profiles (GI). Similarly, Szakacs [27] and co-workers correlated expression profiles of all 48 human ABC transporters with patterns of drug activity in the NCI 60 cell lines. They were able to identify candidate substrates for several ABC transporters and compounds whose toxicities are potentiated by ABCB1-MDR1.

One potential application of microarrays in toxicology is their use in predicting toxicity of undefined chemicals by comparing their gene expression patterns in a biological model with databases of microarray-generated gene expression data corresponding to known toxicants. Feasibility of compound classification based on gene expression profiles is proven by several experiments. Hammad *et al.* (2002a, 2002b), for example, analyzed rat liver gene expression patterns elicited by peroxisome proliferators, and enzyme inducers. These authors used several computational analyses including hierarchical clustering [6], PCA, pairwise Pearson correlation of gene

expression profiles, and finally a combination of a genetic algorithm and K-nearest neighbor (GA/KNN) [28]. Their results confirm that compound classification based on gene expression is feasible, and showed both strong within-class correlation of expression profiles and between-class highly distinguishable patterns.

The work of Staunton *et al.* [29] is an example of a supervised-learning approach. Specifically, the authors investigated whether patterns of gene expression were sufficient to predict sensitivity or resistance of the NCI 60 cell lines to 232 chemical compounds whose GI activity profile had been previously measured. They measure gene expression of 6817 genes in each of the 60 untreated cell lines using Affymetrix chips. Chemosensitivity prediction was modeled as a binary classification problem, and thus for each compound two classes of cell lines were defined: sensitive (class 1) and resistant (class 2), according to the GI profiles. They then divided the data set into a training set and a test set. The classifier was implemented using a weighted voting algorithm, in which correlated genes "vote" on whether a cell is predicted to be sensitive or resistant. Correlation in the training set between a compound  $c$  and a gene  $g$  is defined as:

$$P(g,c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(d) + \sigma_2(d)} \quad (2)$$

Large values of the correlation  $P(g,c)$  indicate that the gene expression is a good indicator of class distinction. A weighted sum of the gene expression level of strongly correlated genes is then used to classify. Classifiers with up to 200 genes were tested, with the median accuracy of the classifiers reaching 75%. From this work one can conclude that indeed gene expression profiles in untreated cells can be used to predict whether a cell line is sensitive or resistant to a particular drug.

Other interesting examples of supervised classification methods applied to drug-treated human neural cell cultures come from two studies of Gunther and colleagues. The aim of first study [30] was to investigate whether high content statistical categorization of drug-induced gene expression profiles can be used to predict the drug's therapeutical class among different classes of psychoactive compounds. Primary cultures of human neuronal precursor cells were treated with multiple members of antidepressants (AD), antipsychotics (AP), and opioid receptor agonists (OP). Arguably, however, one of the most used class of psychoactive drugs, the class of antianxiety compounds, would have been an interesting choice. Gene expression was measured using DNA microarrays containing about 11k oligonucleotide probes. Data was analyzed by supervised statistical classification including Classification Tree (CT) and Random Forest (RF) methods. Both methods are based on a "leave-one-out" training and testing series, so that the class of the naive test sample can be predicted after training over all other samples. The former method resulted in 88.9% of correct predictions, and relied on few strong markers. Notably, accuracy did not decline significantly when the classification was repeated after withholding the predominant classifier genes from the analysis. The latter method is based on stochastic feature evaluation, and resulted in a correct prediction rate of 83.3% relying on a

much larger set (326) of weak marker genes. Interestingly, two examples are given in which one subclass of AD (SSRIs, or tricyclic) could be successfully predicted as belonging to the antidepressants class after being excluded from the training using the RF. Although the accuracy of prediction of novel subclass unrepresented in the training was surprisingly high (100%), it is unclear why a similar analysis after withholding the third subclass of AD adopted in this study, the MAOIs, is missing. The authors of this work recently published a new study [31] in which they propose a novel algorithm for drug efficacy-profiling, called Sampling Over Gene Space (SOGS), and applied it to drug-treated human cortical neuron 1A cell line. While less appealing from a physiological point of view, cell line monocultures provide a simpler system more suitable for reproducible chemical genomics screening. This procedure is based on supervised classification methods such as Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM), expected to yield stronger predictions than stochastic feature evaluation such as RF on one hand, but on the other they are more prone to ‘overfitting’ the training data. SOGS however builds multiple classifier methods iteratively sampling random sets of features using LDA or SVM, and the final classification is based on the most frequent classification over the multiple iterations. The authors claim that such a combination of stochastic feature evaluation with the stable LDA and SVM modeling methods minimizes overfit, while increasing prediction strength.

### 3. NETWORK/PATHWAY RECONSTRUCTION

Perturbations to the state of the cell have been used extensively in molecular biology to infer the function of a single gene or protein. With the advent of high throughput quantitative methods it has become possible to move from a qualitative biology to a quantitative biology, thus enabling the use of methodologies typical of engineering and physics to the study of the biological processes and the emergence of “systems biology”, *i.e.* the integrated study of biological processes (for a good review of systems biology refer to Brent [32] and for its application in drug discovery refer to Butcher *et al.* [33] and Apic *et al.* [34]). Biological processes are the result of complex interaction among thousands of components. Network, or, graph theory, is a mathematical formalism that is very well suited for describing such interactions. Hence the renewed interest in network theory and its potential impact on molecular biology and medicine.

In the area of drug development, particular relevance assumes “reverse engineering” whose goal is to map gene, protein and metabolite interactions in the cell, thus elucidating the regulatory circuits used by the cell for its functioning, and their malfunctioning during diseases. A very good review was recently published on this topic [35].

We can distinguish two different reverse engineering strategies [35]: the “physical approach” and the “influence approach”. In the former, the aim is to use RNA expression data to identify the transcription factors (TFs) and the DNA binding sites to which the factor binds. The interactions thus inferred are true physical interactions between TFs and the promoters of the regulated genes. In the latter, the aim is to find regulatory influences between RNA transcripts that do not necessarily have to be of the TF-DNA binding site kind.

The general model, as shown in Fig. 2, requires that some RNA transcripts act as regulatory “inputs” whose concentration variations drive the expression of an “output” transcript. Such a model therefore does not describe physical interactions, since an mRNA does not control directly the level of other mRNAs, but rather aims at inferring the regulatory influence between two or more transcripts that may as well be indirect through the action of proteins, metabolites and other molecules.

Reverse engineering algorithms make use of measurements of transcript concentrations in response to perturbations to the state of the cell in order to infer regulatory interactions.

#### 3.1. Target Identification and Validation

For a detailed description on computational and bioinformatics methods to infer interactions among genes and proteins we refer the interested reader to an excellent review on this topic by Gardner *et al.* [35]. Here we will briefly discuss two recent examples based on two different methodologies that use the “influence strategy” as defined above.

The first methodology describes a gene network as a system of ordinary differential equations [36]. The rate of change in concentration of a particular transcript,  $x_i$ , is given by a nonlinear influence function,  $f_i$ , of the concentrations of other RNAs:

$$\frac{dx_i}{dt} = f_i(x_1, \dots, x_n) \quad (3)$$

Where  $n$  is the number of genes or transcripts in the network. The function  $f_i$  can have different forms. The easiest form that this function can assume is the linear form where Equation (3) becomes:

$$\frac{dx_i}{dt} = \sum_j w_{ij} x_j + p_i \quad (4)$$

where  $w_{ij}$  represents the influence of gene  $j$  on gene  $i$ , and  $p_i$  an externally applied perturbation to the level of transcript  $i$ . We developed an inference algorithm named Network Identification by Regression (NIR) [37] that uses the differential equation model of a gene network in Equation (4) to infer the regulatory interactions among 9 genes part of the *Escherichia coli* SOS pathway. The strategy we adopted was to overexpress each of the 9 genes in the network using an exogenous plasmid carrying a copy of the gene under the control of an inducible promoter. After transfection and induction of the vector, the gene expression change of the 9 genes in the network was measured at steady-state, *i.e.* when the cell has reached a new equilibrium and all the transient effects are over. Under these conditions, the term on the left hand-side of Equation (4) becomes  $\frac{dx_i}{dt} = 0$ , so that the

equation can be rewritten as:

$$-p_i = \sum_j w_{ij} x_j \quad (5)$$

where both  $p_i$  and  $x_j$  for all the 9 different perturbation experiments are experimentally measured, whereas the

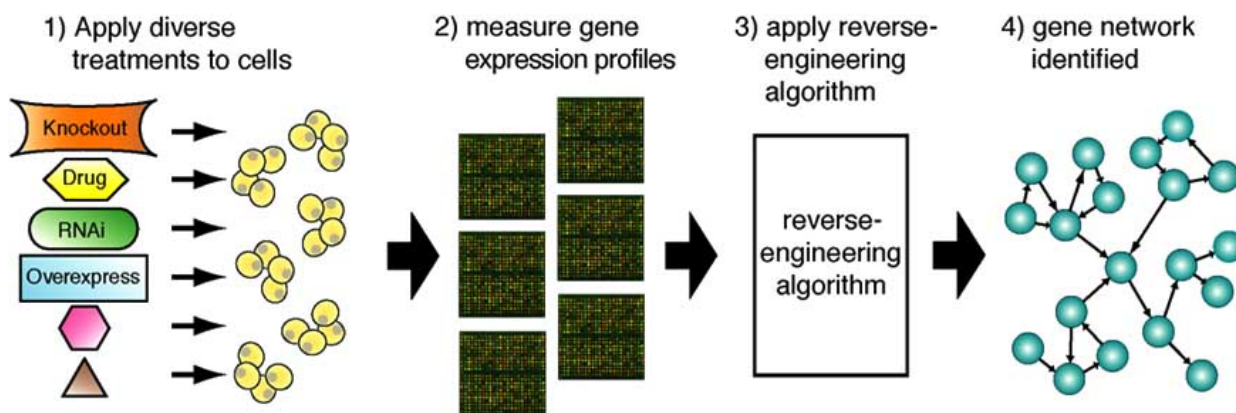


Fig. (2). Schematic diagram of reverse-engineering approaches to drug discovery. Gene expression profiles following a variety of perturbations to the cells are used to reconstruct the network of interactions of gene, proteins and metabolites.

weights  $w_{ij}$  are the unknown parameters that we would like to learn from the data. Using multiple linear ridge regression, we were able to recover a network model, shown in Fig. 3, that correctly identified 25 of the previously known regulatory interactions between the 9 transcripts, as well as 14 interactions that could be novel, or possibly false positives. These results were obtained with a noise-to-signal ratio of 68%. From a drug discovery point of view, this approach would be powerful for finding new targets for antibiotics, since the 9 genes are part of the SOS pathway involved in response to DNA damage. The genes that are the 'hubs' of the network, *i.e.* those genes that are the main regulators of the system, are ideal targets for new antibiotics because they would block the response of the bacteria to damage, thus preventing their survival.

As a second example of successful network inference applied to a mammalian system, we will illustrate the work of Basso *et al.* [38]. The approach used by these authors is based on information theory. Their approach named ARACNE is based on the computation of mutual information among pair of genes. For a pair of discrete random variables,  $x$  and  $y$ , the mutual information is defined as

$$I(x, y) = S(x) + S(y) - S(x, y) \quad (6)$$

where  $S(\cdot)$  defines the entropy. For a given discrete stochastic variable  $t$  the entropy is defined as:

$$S(t) = \sum_i \Pr(t = t_i) \log(\Pr(t = t_i)) \quad (7)$$

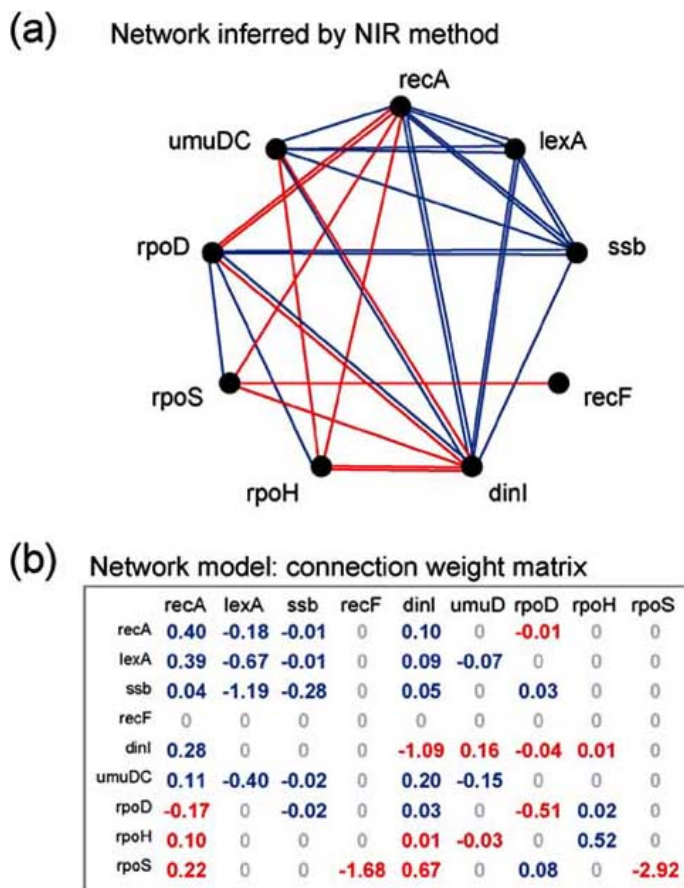
As it can be intuitively appreciated from the above definition entropy is maximal for a uniformly distributed variable. The probability is estimated using Montecarlo simulations. To each value of the mutual information  $I(x, y)$  is associated a  $p$ -value computed again using Montecarlo simulations. The null hypothesis associated to the  $p$ -value corresponds to pair of nodes that are disconnected from the network and from each other. The final step of their algorithm is a pruning step that tries to reduce the number of false positives (*i.e.* inferred interactions among two genes that are not direct interaction in the real biological pathway). They use Data Processing Inequality principle that asserts that if both  $(x, y)$  and  $(y, z)$  are directly interacting, and  $(x, z)$  are indirectly interacting through  $y$ , then  $I(x, z) \leq I(x, y)$  and

$I(y, z) \leq I(x, y)$ . This condition is sufficient but not necessary, *i.e.* the inequality can be satisfied even if  $(x, z)$  are directly interacting, therefore the authors acknowledge that by applying this pruning step using DPI they may be discarding some direct interactions as well. The authors applied their algorithm on a data set consisting of 336 whole-genome expression profiles representative of perturbations of B cell lines and are able to find novel direct targets of the Transcription Factor MYC.

### 3.2. Hit Identification, Lead Identification and Optimization

Network identification can be used to infer the direct gene and protein targets of a compound with unknown mode of action. One of the earliest approaches of this kind has been proposed by Imoto *et al.* [39]. Although the approach described in the paper is somewhat confusing, we decided to include it in our review since to our knowledge this is one of the first papers to propose that network inference can be used for lead optimization. The authors termed their approach the "virtual gene technique". Briefly, using an algorithm by Maki *et al.* [40] they reconstruct a directed acyclic graph (DAG) describing gene regulatory interactions considering the drug as a "virtual gene". Let  $V = \{g_1, g_2, \dots, g_n\}$  the set of all genes and  $D = \{d_1, d_2, \dots, d_m\}$  the set of genes to be knocked out in order to perturb the system.  $D$  is assumed to contain also the virtual gene and the perturbation experiment associated to this virtual gene is treatment with the drug. By observing how the genes change in response to the gene disruption they are able to find a DAG by drawing an edge between two nodes of the graph if a certain equivalence relationship is satisfied. By considering the DAG whose root is the virtual gene, the children of this virtual gene would be the candidate genes directly affected by the drug. From their paper is not clear how well their method performs since the experimental results on deletion strains of *S. cerevisiae* are poorly described. However, their method is an illustrative example of how network inference can be applied to drug discovery.

Another example of network inference to drug discovery is the work of Haggarty *et al.* [41]. Their approach is based on wildtype and nine different gene deletion strains in *S. cerevisiae*. Each of the strains is treated with all the possible combinations of 2 molecules drawn from a set of 24 small



**Fig. (3).** Inference of a nine-transcript subnetwork of the SOS pathway in *E. coli* using the NIR algorithm. (a) Graph depiction of the network model identified by the NIR algorithm. Previously known regulatory influences are marked in blue, novel influences (or false positives) are marked in red. The strengths and directions of the identified connections are not labeled in the graph. (b) The network model is also depicted as a matrix of interaction strengths. The colors are the same as in panel (a).

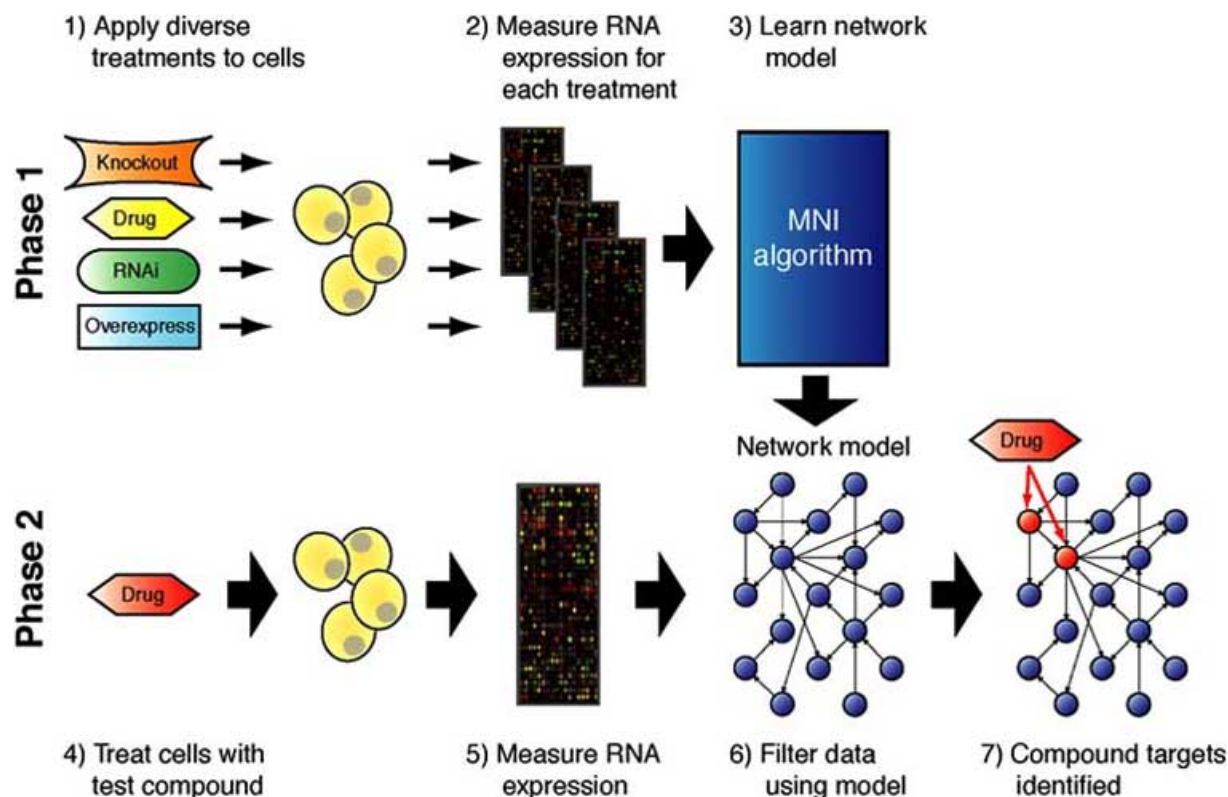
molecules. The authors propose a method that can be used to understand which of the molecules have similar mode of action by measuring the similarity of chemo-genomic networks. For each strain the data were represented as an adjacency matrix,  $A$ , with one row and one column for each of the 24 molecules tested. The element  $a_{ij}$  of matrix  $A$  is 0 when no observable effect on growth after treatment with compound  $i$  and  $j$  is found, 1 if there is a measurable growth defect. For each compound in each strain, information in  $A$  can be used for clustering the compounds on the basis of similarity in their pattern of biological activity. However the authors do not test thoroughly this prediction.

The NIR algorithm we developed and briefly described in section 3.1, can also be used for compound mode of action discovery. The network model can be used as a predictive tool for analyzing new RNA expression data obtained by measuring transcript responses to a drug treatment. As a proof-of-principle, we applied the antibiotic mitomycin C to *E. coli* and observed changes in all nine measured SOS transcripts. However, the known mediator of mitomycin C is only the gene *recA*. The network model obtained by the NIR algorithm enables us to separate secondary changes from primary changes due to direct interaction with the drug. In this case Equation (5) can be solved to find the  $p_i$  value for

each  $i = 1 \dots 9$ , since the network model  $w_{ij}$  is known while  $x_j$  are the measured response of the cell to the drug treatment. If  $p_i$  is close to 0, then gene  $i$  is not a direct target of the drug, otherwise gene  $i$  is directly interacting with the compound.

The network model correctly filters the RNA expression response to the drug treatment to reveal the *recA* gene as the direct target. The same target was identified for treatment with UV irradiation and the antibiotic pefloxacin, both of which stimulate *recA* transcript, but not for novobiocin, a drug that should not directly interact *via* the *recA* gene.

We recently proposed an extension of the NIR algorithm called Mode of action by Network Identification (MNI), that computes the likelihood that gene products and associated pathways are targets of a compound [42]. Our approach is described in Fig. 4. We first reverse-engineer a network model of regulatory interactions in the organism of interest using a training data set of whole-genome expression profiles. The network model is based on ordinary linear differential equations under steady-state conditions described by Equation (5). We then use the model to analyze the expression profile of the compound-treated cells to determine the pathways and genes targeted by the compound. The algorithm assumes that the expression profile training data set are obtained at steady-state following



**Fig. (4).** Overview of the NMI method. In phase 1, a set of treatments is applied to cells. Changes in mRNA species are measured. The data are then used by the MNI algorithm to infer a model of the regulatory network among the genes. In phase 2, cells are treated with the test compounds and the expression changes of all the mRNA species is measured. The expression data are then filtered using the network model to distinguish the targets of the test compound from secondary responders.

a variety of treatment, including compounds, RNAi, and gene-specific mutations (Fig. 4).

The ability to use different treatment types is an important advance over earlier model estimation techniques that require knowledge of the targets of the perturbations. To infer a network model without requiring gene-specific perturbations the algorithm employs an iterative procedure. It first predicts the targets of treatment using an assumed network model, and then uses those predicted targets to estimate a better model. The procedure stops once convergence criteria are met. Once the regulatory model has been learned, we applied it to the expression profile of a test compound to predict its targets. We applied this method to the *S. cerevisiae* using as a training data set 515 whole-genome yeast expression profiles resulting from a variety of treatment [9,43]. We then used MNI algorithm to identify the probable targets of 15 compounds, 13 of which were drawn from the Hughes compendium [9] and from other studies [44]. Of these 15 compounds, 9 had previously known targets, while the targets of other six were previously unknown. MNI ranks the ~6000 genes in yeast according to their probability of being direct targets of the compound. By selecting the top 50 genes predicted by MNI for a compound, it is possible to infer the pathways directly affected by the drug looking for significantly overrepresented Gene Ontology (GO) processes among the highly ranked genes.

For 7 out of 9 compounds with known MOA, MNI correctly identified the known target pathway and for 6 out of this 9 it was able also to identify the correct target gene. We then demonstrated the use of MNI on a tetrazole-containing compound, 1-phenyl-1H-tetrazole-5-ylsulfonyl-butanenitrile (PTSB) found to inhibit both wt *S. cerevisiae* and human small lung carcinoma cells. We applied MNI to the expression profile after treatment with PTSB and found two genes: thioredoxin reductase (TRR1, MNI\_rank=32) and thioredoxin (TRX2, MNI\_rank=36) while the overrepresented GO process among the top 50 genes was the 'cell redox homeostasis'. We validated the prediction made by MNI with appropriate biochemical assays and confirmed that PTSB acts by inhibiting these two targets.

#### 4. DISCUSSION

Computational biology and bioinformatics approaches have the potential to completely change the way drugs are discovered and designed. Already these methods are having an impact on the different stages of the drug discovery process. We have shown in this review how computational methods like classification and network-based algorithms can be used to understand the mode of action and the efficacy of a given compound and to help elucidating the pathophysiology of a disease. But these computational tools, in our opinion, may also be used in a different and innovative way to promote a change of paradigm in how drugs are designed. In the pharmacological industry there

has already been a shift from symptomatic oriented drugs, that can relieve the symptoms but not the cause of the disease, to pathology-based drugs whose targets are the genes and proteins involved in the etiology of the disease. Drugs targeting the affected pathway have thus the potential to become therapeutic. An example of this is enzyme replacement therapy in genetic sulfatase deficiency syndromes [45]. The sulfatase enzyme is the missing protein that when reintroduced in the organism is able to restore the pathway that had been altered by the disease process. The passage from symptomatic-centered drug discovery to disease-centered drug-discovery has been forced upon the industry by the availability of the full sequence of the human genome, with its implicit promise of novel potential targets. However, as reported in a recent review by Csermely *et al.* [46], the number of successful drugs did not increase appreciably in the recent years. With the current paradigm, an ideal drug is both potent and specific, *i.e.* it targets specifically a single protein. In our opinion a second shift is now necessary and will be driven by the availability of sophisticated computational biology and bioinformatics tools: a shift from single-target drugs to “network drugs”. By network drug we define a compound or a set of compounds that is able to alter a biological pathway disregulated by a disease in a predefined way so as to restore its normal physiological function. A similar concept has been put forward by Csermely *et al.* [46] in their review, where they propose the partial inactivation of multiple targets as a novel paradigm for drug design. They argue that such kind of multi-target drug could be much more efficient than a drug directed at a single target. They proposed that a network approach to drug design would examine the effect of drugs in the context of a network of relevant protein-protein, regulatory and metabolic interactions. The end result would be the development of a drug that would hit multiple targets selected in such a way as to decrease network integrity and so completely disrupt the functioning of the network. Our idea is to take this approach one step further and aim not at disrupting the network, but into developing compounds and delivery techniques able to change the behavior of the network in a controllable and predictable manner.

Thanks to network-inference approaches, some of which were described in this review, it is now becoming possible to have a detailed map of the regulatory circuit among genes, proteins and metabolites. This in turn allows a better understanding of how biological pathways are regulated and how they accomplish their function. The approaches presented in this review also allow the screening of a compound to quickly identify the proteins it interacts with. This gives us all the necessary tools to identify and repair the disregulated biological pathway causing the disease, much as an engineer would do to restore a malfunctioning electronic circuit. If she/he finds that a specific component of the circuit is malfunctioning, it would be bypassed using extra wires that would bridge different parts of the circuit. Sometimes this would not be sufficient since those parts of the circuit should be in contact only under precisely defined conditions. In this case, she/he would also need to add a microchip that would take care of activating those connections only when necessary.

Similarly one could think of delivering multiple compounds, each directed to a specific biological target, in a

coordinated way controlled by a computer chip that would release the drugs in the organism only when needed to restore physiological behavior of the pathway disregulated by the disease. The key step in this approach is to have a detailed knowledge of the network of protein, gene and metabolite interactions in the different biological pathways.

Although this picture may seem farfetched all the tools to accomplish this feat have already been developed and are here to stay, and hopefully in the next decades the way we think of drugs will be completely different.

## REFERENCES

- [1] DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ* **2003**; *22*: 151-85.
- [2] Ratti E, Trist D. The continuing evolution of the drug discovery process in the pharmaceutical industry. *Farmaco* **2001**; *56*: 13-9.
- [3] Hart CP. Finding the target after screening the phenotype. *Drug Discov Today*, **2005**; *10*: 513-9.
- [4] Bredel M, Jacoby E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* **2004**; *5*: 262-75.
- [5] Fagan R, Swindells M. Bioinformatics, target discovery and the pharmaceutical/biotechnology industry. *Curr Opin Mol Ther* **2000**; *2*: 655-61.
- [6] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **1998**; *95*: 14863-8.
- [7] Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*, Springer, New York, NY 2001.
- [8] Stoughton RB, Friend SH. How molecular profiling could revolutionize drug discovery. *Nat Rev Drug Discov* **2005**; *4*: 345-50.
- [9] Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **2000**; *296*: 1205-14.
- [10] Hartigan JA. *Clustering Algorithms*, John Wiley & Sons, New York, NY 1975.
- [11] Gasch AP, Spellman PT, Kao CM, *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **2000**; *11*: 4241-57.
- [12] Brown MP, Grundy WN, Lin D, *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* **2000**; *97*: 262-7.
- [13] Stegmaier K, Ross KN, Colavito SA, O'Malley S, Stockwell BR, Golub TR. Gene expression-based high-throughput screening (GHTS) and application to leukemia differentiation. *Nat Genet* **2004**; *36*: 257-63.
- [14] Paull KD, Shoemaker RH, Hodes L, *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J Natl Cancer Inst* **1989**; *81*: 1088-92.
- [15] Weinstein JN, Myers TG, O'Connor PM, *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**; *275*: 343-9.
- [16] Bao L, Guo T, Sun Z. Mining functional relationships in feature subspaces from gene expression profiles and drug activity profiles. *FEBS Lett* **2002**; *516*: 113-8.
- [17] Marton MJ, DeRisi JL, Bennett HA, *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* **1998**; *4*: 1293-301.
- [18] Parsons AB, Brost RL, Ding H, *et al.* Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotechnol* **2004**; *22*: 62-9.
- [19] Giaever G, Shoemaker DD, Jones TW, *et al.* Genomic profiling of drug sensitivities *via* induced haploinsufficiency. *Nat Genet* **1999**; *21*: 278-83.
- [20] Giaever G, Flaherty P, Kumm J, *et al.* Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci USA* **2004**; *101*: 793-8.
- [21] Baetz K, McHardy L, Gable K, *et al.* Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action. *Proc Natl Acad Sci USA* **2004**; *101*: 4525-30.

- [22] Lum PY, Armour CD, Stepaniants SB, *et al.* Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell* **2004**; 116: 121-37.
- [23] Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. *Science* **2004**; 306: 1194-8.
- [24] Betts JC, McLaren A, Lennon MG, *et al.* Signature gene expression profiles discriminate between isoniazid-, thiolactomycin-, and triclosan-treated *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **2003**; 47: 2903-13.
- [25] Scherf U, Ross DT, Waltham M, *et al.* A gene expression database for the molecular pharmacology of cancer. *Nat Genet* **2000**; 24: 236-44.
- [26] Dan S, Tsunoda T, Kitahara O, *et al.* An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res* **2002**; 62: 1139-47.
- [27] Szakacs G, Annereau JP, Lababidi S, *et al.* Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. *Cancer Cell* **2004**; 6: 129-37.
- [28] Li L, Darden TA, Weinberg CR, Levine AJ, Pedersen LG. Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Comb Chem High Throughput Screen* **2001**; 4: 727-39.
- [29] Staunton JE, Slonim DK, Collier HA, *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA* **2001**; 98: 10787-92.
- [30] Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles *in vitro*. *Proc Natl Acad Sci USA* **2003**; 100: 9608-13.
- [31] Gunther EC, Stone DJ, Rothberg JM, Gerwien RW. A quantitative genomic expression analysis platform for multiplexed *in vitro* prediction of drug action. *Pharmacogenomics J* **2005**; 5: 126-34.
- [32] Brent R. A partnership between biology and engineering. *Nat Biotechnol* **2004**; 22: 1211-4.
- [33] Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. *Nat Biotechnol* **2004**; 22: 1253-9.
- [34] Apic G, Ignjatovic T, Boyer S, Russell RB. Illuminating drug discovery with biological pathways. *FEBS Lett* **2005**; 579: 1872-7.
- [35] Gardner TS, Faith JJ. Reverse-engineering transcription control networks. *Phys Life Rev* **2005**; 2: 65-88.
- [36] De Jong H, Gouze JL, Hernandez C, Page M, Sari T, Geiselmann J. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol* **2004**; 66: 301-40.
- [37] Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action *via* expression profiling. *Science* **2003**; 301: 102-5.
- [38] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet* **2005**; 37: 382-90.
- [39] Imoto S, Savoie CJ, Aburatani S, *et al.* Use of gene networks for identifying and validating drug targets. *J Bioinform Comput Biol* **2003**; 1: 459-74.
- [40] Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y. Development of a system for the inference of large scale genetic networks. *Pac Symp Biocomput* **2001**: 446-58.
- [41] Haggarty SJ, Clemons PA, Schreiber SL. Chemical genomic profiling of biological networks using graph theory and combinations of small molecule perturbations. *J Am Chem Soc* **2003**; 125: 10543-5.
- [42] Di Bernardo D, Thompson MJ, Gardner TS, *et al.* Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol* **2005**; 23: 377-83.
- [43] Mnaimneh S, Davierwala AP, Haynes J, *et al.* Exploration of essential gene functions *via* titratable promoter alleles. *Cell* **2004**; 118: 31-44.
- [44] Ueda M, Kinoshita H, Yoshida T, Kamasawa N, Osumi M, Tanaka A. Effect of catalase-specific inhibitor 3-amino-1,2,4-triazole on yeast peroxisomal catalase *in vivo*. *FEMS Microbiol Lett* **2003**; 219: 93-8.
- [45] Cosma MP, Pepe S, Annunziata I, *et al.* The multiple sulfatase deficiency gene encodes an essential and limiting factor for the activity of sulfatases. *Cell* **2003**; 113: 445-56.
- [46] Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol Sci* **2005**; 26: 178-82.
- [47] Walker MG. Pharmaceutical target identification by gene expression analysis. *Mini Rev Med Chem* **2001**; 1: 197-205.
- [48] Parsons AB, Geyer R, Hughes TR, Boone C. Yeast genomics and proteomics in drug discovery and target validation. *Prog Cell Cycle Res* **2003**; 5: 159-66.
- [49] Bugrim A, Nikolskaya T, Nikolsky Y. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov Today* **2004**; 9: 127-35.
- [50] Hamadeh HK, Bushel PR, Jayadev S, *et al.* Prediction of compound signature using high density gene expression profiling. *Toxicol Sci* **2002**; 67: 232-40.
- [51] Hamadeh HK, Bushel PR, Jayadev S, *et al.* Gene expression analysis reveals chemical-specific profiles. *Toxicol Sci* **2002**; 67: 219-31.