Expression profiling of drug response—from genes to pathways

Ralf Herwig, PhD; Hans Lehrach, PhD

Recent reports have highlighted the imbalance between rising costs in drug discovery and the production of new molecular entities for the market,[1,2] leading to a long-term loss of efficiency. Remarkably, this decline in productivity has occurred despite the fact that biomedical research benefits from large governmental and private investments, and despite the comprehensive improvements in our knowledge of human genes resulting from large sequencing projects.

The tremendous efforts that have to be invested for drug target identification, follow-up validation studies, and clinical trials, in combination with the high failure rate as a consequence of individual response to drugs, has imposed high costs on the development of drugs. Understanding individual response to a drug, what determines its efficacy and tolerability in the patient's body, is the major bottleneck in drug development and

Understanding individual response to a drug—what determines its efficacy and tolerability—is the major bottleneck in current drug development and clinical trials. Intracellular response and metabolism, for example through cytochrome P-450 enzymes, may either enhance or decrease the effect of different drugs, dependent on the genetic variant. Microarrays offer the potential to screen the genetic composition of the individual patient. However, experiments are "noisy" and must be accompanied by solid and robust data analysis. Furthermore, recent research aims at the combination of high-throughput data with methods of mathematical modeling, enabling problem-oriented assistance in the drug discovery process. This article will discuss state-of-the-art DNA array technology platforms and the basic elements of data analysis and bioinformatics research in drug discovery. Enhancing single-gene analysis, we will present a new method for interpreting gene expression changes in the context of entire pathways. Furthermore, we will introduce the concept of systems biology as a new paradigm for drug development and highlight our recent research—the development of a modeling and simulation platform for biomedical applications. We discuss the potentials of systems biology for modeling the drug response of the individual patient.

Dialogues Clin Neurosci. 2006;8:283-293.

Keywords: *drug discovery; functional genomics; microarray; bioinformatics; data integration; database; systems biology*

Author affiliations: Max Planck Institute for Molecular Genetics, Department of Vertebrate Genomics, Berlin, Germany

Address for correspondence: Dr Ralf Herwig, Max Planck Institute for Molecular Genetics, Department of Vertebrate Genomics, Ihnestr. 73, D-14195 Berlin, Germany
(e-mail: herwig@molgen.mpg.de)

Selected abbreviations and acronyms

AD	*Alzheimer's disease*
ALS	*amyotrophic lateral sclerosis*
DRPLA	*dentatorubral-pallidoluysian atrophy*
GEO	*gene expression omnibus*
GO	*gene ontology*
GPCR	*G-protein-coupled receptor*
HD	*Huntington's disease*
PCR	*polymerase chain reaction*
PD	*Parkinson's disease*
SAGE	*serial analysis of gene expression*
SOP	*standard operating procedure*

clinical trials. When a drug is delivered through the body, each individual reacts differently in terms of intracellular response and metabolism. A prominent example is seen with the cytochrome P-450 enzymes, a family of drug-metabolizing enzymes that may either enhance or decrease the effect of different drugs, dependent on the genetic variant.[3] Thus, the individual genetic composition of the patient has become a major issue in studying drug targets and responses to medical treatment.

Microarrays are the state-of-the-art platform for screening the genetic composition of the individual patient. This technology offers the chance to acquire the complete state of gene expression[4-6] and to identify genes and pathways that are affected by the treatment.[7,8] On the other hand, high-throughput technologies such as microarrays are also a part of the problem. The new technologies have led to an increasing amount of heterogeneous (and often conflicting) data, corresponding to an increasing amount of potential drug targets.

Microarray experiments are "noisy" by nature, and must be accompanied by solid and robust data analysis components. This task has been part of bioinformatics research since the advent of this new discipline. The components of microarray analysis range from low-level analysis, explorative statistics to higher-level analysis involving additional data, annotation, and knowledge in order to embed the gene expression data in a functional context. The main purpose of data analysis is to filter the information and to enrich the level of information complexity from single gene markers to biological pathways.

This article will discuss the state-of-the-art deoxyribonucleic acid (DNA) array technology platforms and the basic elements of data analysis and bioinformatics research in drug discovery, developed by us and others. Apart from the single-gene analysis we will present a new method for interpreting gene expression changes in the context of the pathways involved. Recent microarray applications for neuroscience will be considered, and the particular challenges for gene expression analysis of the brain will be discussed. Furthermore, we will introduce the concept of systems biology as a new paradigm for drug development and highlight our recent research—the development of a modeling and simulation platform for biomedical applications. This research field, which shows great potential for modeling the drug response of the individual patient, will deliver valuable hypotheses for personalized drug treatment and therapy monitoring in the medium to long term.

DNA array platforms for gene expression profiling

DNA arrays are the most common gene expression profiling technology. A DNA array consists of a solid support (nylon membrane, glass chip) that carries DNA sequences representing genes—the probes. In hybridization experiments with the target sample of labeled complementary ribonucleic acids (cRNAs) and through subsequent data capture a numerical value, the signal intensity, is assigned to each probe. Labeling is done either radioactively (phosphorus, ^{33}P) and detected with a phosphor imager or fluorescently (Cy3/Cy5 dyes) and detected with specific scanners. Chips are typically small (<2 cm^2) and allow the immobilization of tens of thousands of different gene representatives.

The most prominent DNA array technology is the Affymetrix GeneChip system.[9] Here, genes are represented by probe sets of short oligonucleotides (typically 11 to 20 25mers) that are distributed across their sequences. These oligonucleotides are synthesized in a highly specific manner at defined locations using a photolithographic procedure. After hybridization, the measured intensity for the represented gene is summarized across the different probes in the probe set. Affymetrix chips have emerged as the pharmaceutical standard, and are widely in use because of the highly standardized chip generation process. Whole-genome chips are available for a large number of organisms, such as human, mouse, rat, bovine, pig, etc. An experiment with Affymetrix technology is typically a single-channel experiment, ie, only one target sample is analyzed in one experiment.

An alternative technology is the Agilent system.[10] This relies on the immobilization of longer oligonucleotides (60mers) synthesized in situ at or near the surface of the slide by inkjet printing using phosphoramidite chemistry. These probes are highly specific for the represented gene and show, generally, better hybridization properties than systems with shorter oligonculeotides. Experiments are typically double-channel experiments, ie, two target samples are analyzed simultaneously, each labeled with a different cyanine dye and quantified with a separate scanning procedure.

A recent technological development is the Illumina BeadChip system[11,12] that utilizes an "array of arrays" format. Each array on the support contains thousands of wells into which up to hundreds of thousands of beads self-assemble in a random fashion. Specific 50-mer gene sequences concatenated with an address sequence recognize the beads and attach to them. After bead assembly, a hybridization-based procedure is used to map the array, to determine which bead type resides in each well of the array and to validate the performance of each bead type. An advantage of this technology is that several samples can be analyzed on the same chip, thus preventing experimental artifacts across chips or dye labeling procedures. For example, the recent HumanRef-8 chip offers the possibility of screening eight different samples in parallel.

Other commercial chip providers are Amersham Biosciences, NimbleGen, Febit, and Applied Biosystems. There are advantages and disadvantages of the above-mentioned platforms regarding hybridization specificity, sample target material needed, and other factors, as pointed out in a recent review.[13]

Historically, the first array technology was based on spotted cDNAs.[14-16] This technology is still extensively in use in the academic sector, but also in pharmaceutical research

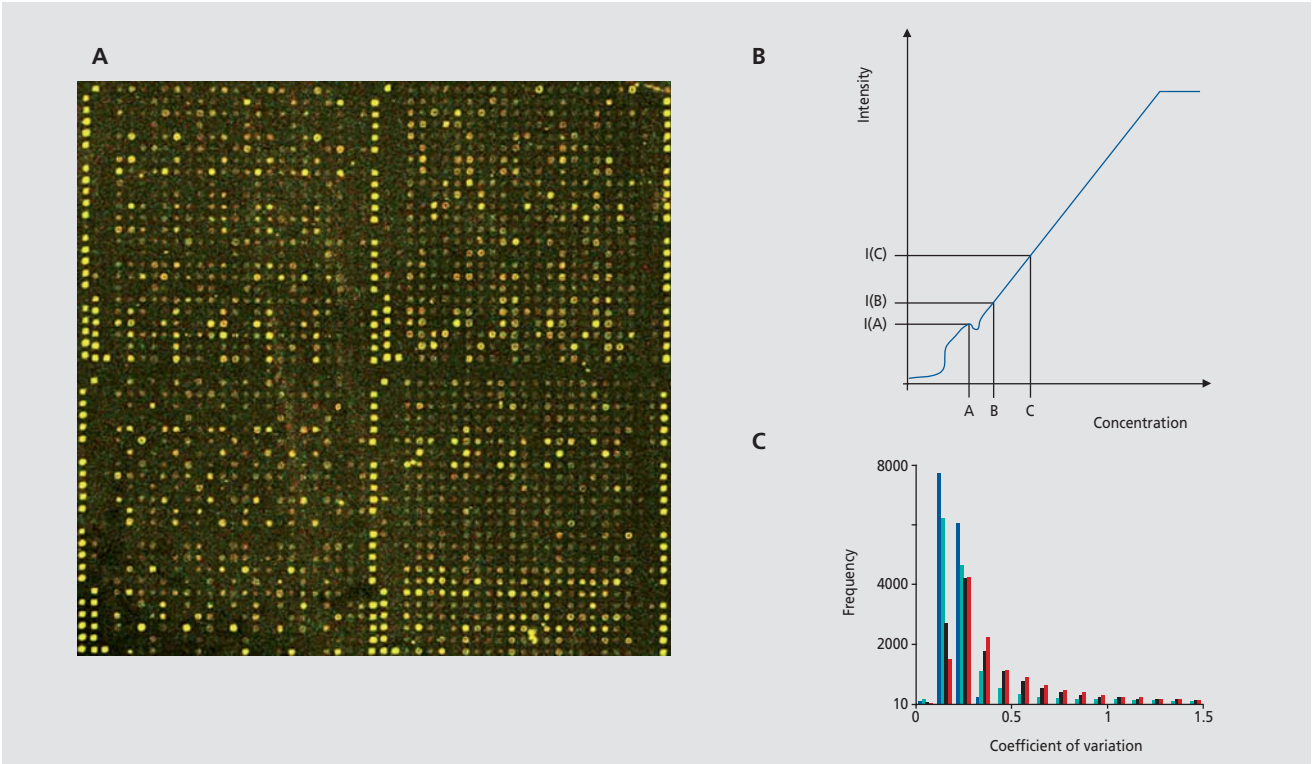

Figure 1. **A**: False-color image generated from a two-color hybridization on a cDNA array.[17] **B**: Linearity between concentration and measured signal intensity is the underlying assumption of microarray data analysis. Whereas the expression ratio of genes B and C yield a valid measure of the concentration differences, the ratio of genes A and B is misleading because of nonlinear deviations in the low intensity region. **C**: Histogram of the coefficient of variation for genes from simulated array images[26] using three different image analysis programs for data analysis that can be classified as manual (red), semi-automatic (black) and fully automatic (green). The blue bars show the counts for the simulated input data.

that involves probe sets not covered by standard array formats. cDNAs have a high variability in length (600 to 1500 bp) and are amplified using a polymerase chain reaction (PCR). PCR products are then transferred to the surface via contact printing by robotic devices *(Figure 1a)*.

The implicit assumption of all microarray studies is that the signal intensity measured with a specific probe is proportional to the number of molecules of the respective gene in the target sample. Changes in signal intensities are interpreted as concentration changes. It should be pointed out that the signal intensities are only crude estimators for the actual concentrations, and the interpretation as concentration changes is only valid if the intensity-concentration correspondence is approximately linear. Microarray measurements often show deviations from this assumption: for example, saturation effects; the spot signals are above a limit that no longer allows the detection of concentration changes or other nonlinearities if the concentration of the gene is below the detection limit of a microarray *(Figure 1b)*.

Whole-genome chips carry probes for (more or less) the entire genome. These chips are used typically in the beginning of a study when it is not clear what genes are responsible for the drug response of certain groups of patients (for example drug-sensitive and -resistant). For diagnostic purposes specific theme (or custom) chips are used that carry only a few marker genes. The use of custom microarrays for neuroscience applications has been discussed recently.[18]

There have been several studies comparing the performance of microarray platforms.[19-22] Most of these studies reveal a poor correlation in the global expression of the genes. This might be due to several reasons, such as hybridization sensitivity due to the different probe lengths, different chemical treatments, and different statistical methods in the readout of the scanned images. A further issue is the source of the probe sequences. Annotation and probe design typically differ with the background sequence database used by the provider. Currently, several competing collections of transcript sequences are available, and serve as the basis for probe annotation such as Unigene, Refseq, LocusLink, ENSEMBL, etc. Furthermore, probe design of the chip provider must be updated regularly. A recent study showed the potential misinterpretation of experiments performed with Affymetrix probe set assignments that are not updated to the latest genome annotations, and reported a 30% to 50% discrepancy in final lists of differentially expressed genes in several gene expression studies.[23]

Inherent in most technology platforms is software to read the digital image after the scanning process and to compute for each gene representative the intensity value.[24,25] Image analysis methods can be grouped into three different classes: manual, semiautomated, and automated methods. Simulation studies on systematically perturbed artificial images have shown that the data reproducibility increases with the grade of automation of the software *(Figure 1c)*.[26] However, for "noisy" images that show a very irregular structure, manual methods might be the best choice.

Data analysis components

Analysis of expression data comprises several modules that address different questions relevant for drug response screening.[27] The most important tasks are:

- to identify genes that are differentially expressed when comparing two or more conditions (for example, groups of patients resistant or sensitive to a certain drug)
- to identify common gene expression patterns that classify individuals accordingly
- to identify relevant pathways explaining the expression patterns.

Regarding the complexity of the resulting information, the major goal of data analysis is filtering the many thousands of uninformative genes to a set of informative markers, networks, and pathways that are relevant for the problem under analysis *(Figure 2a)*.

Data from microarray experiments typically come out in the form of a table with raw data, ie, the measured intensity values. This raw data is not easily comparable across experimental replicates, so that some *data preprocessing* (or normalization) is necessary. The task of normalization is the elimination of influencing factors that are not due to the probe-target interaction, such as labeling effects (different dyes), background correction, pin effects (spotting characteristics), outlier detection (cross-hybridization of oligonucleotide-probes), etc, thus making signal values comparable across different experiments *(Figure 2b)*. Different algorithms and methods have been proposed to fulfill these tasks.[28-34]

The *identification of differentially expressed genes* between two or more experimental conditions is typically based on two-sample location tests. This setup utilizes replicated experiments with independent samples. The power of such tests is heavily dependent on the number of experimental replicates *(Figure 2c)*. These tests can be used to assign to

each single gene a *P* value that judges the significance of the fold change. Here, it is notable that this *P* value is only valid if the distributional assumptions are valid. For example, if a Student's *t*-test results in a significant *P* value, the implication that the corresponding gene is differentially expressed is only true if both sample series are Gaussian-distributed and have equal variances. Usually, these assumptions do not hold in practice but, strikingly, in most studies this fact is entirely ignored. In our studies we rely therefore on nonparametric alternatives[17,35] *(Figure 2d)*: Wilcoxon's rank sum test is based on the ranks of the replicates rather than on the actual signal values. This test (and other tests based on linear rank statistics such as the van der Waerden test) is preferable to the parametric t-tests if the distributional assumptions cannot be proven to be Gaussian. Furthermore, for "noisy" data this test yields more robust results since it is less sensitive against outlier values. For larger sample sizes, ie, >25 replicates, we can approximate the *P* value of the Wilcoxon rank test by the standard normal distribution. However, most practical applications will be based on a rather smaller number of observations (sample sizes in the order of 4 to 12). Therefore, those *P* values must be calculated exactly. This can be done using a recursive method.[36]

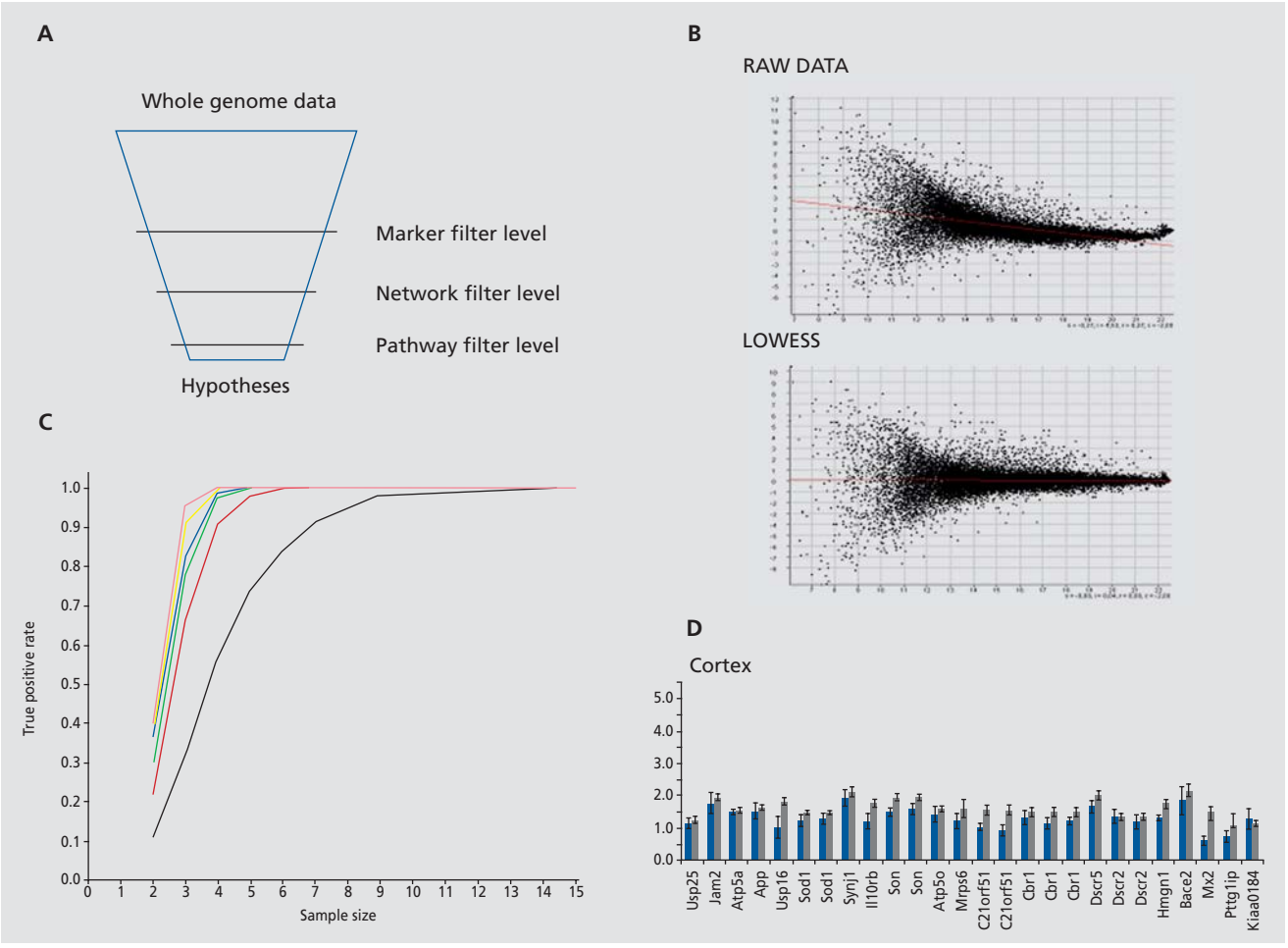

Figure 2. **A**: Schematic description of the biomarker discovery process. **B**: Nonlinear dependencies of fold change (Y-axis) and signal strength (X-axis) in raw data and LOWESS normalization for the compensation of these effects. This method fits the data sets by local polynomials using weighted least squares. **C**: Dependency of detection power for expression differences (Y-axis) from the number of experimental replicates. Different curves correspond to different expression ratios: 1.5 (black), 2 (red), 2.5 (green), 3 (blue), 5 (yellow) and 10 (magenta). **D**: Robust statistical testing identifies even small expression changes (~1.5). Microarray expression changes (gray bars) verified by RT-PCR (red bars) in mouse cortex (kindly provided by Marc Sultan and Marie-Laure Yaspo).[34] LOWESS, locally weighted polynomial regression

If several different experimental conditions are screened (for example different time points after medical treatment), then each gene expresses a certain numerical profile across these conditions. *Clustering algorithms* are explorative statistical methods that group together genes with similar profiles and separate genes with dissimilar profiles, whereby similarity (or dissimilarity) is defined numerically by a pairwise (dis)similarity function such as Euclidean distance or Pearson correlation.[37-40] Hierarchical clustering can be combined with a color-coded representation of the signal values (the expression patterns) and visualized in the form of a dendrogram. Clustering is a very intuitive way of visualizing data, but it should be pointed out that the dendrogram is strongly dependent on the parameters chosen for cluster analysis. Thus, each clustering process should undergo decent validation.[41] Associated groups of genes are usually further investigated, for example for common binding sites in the promoter sequences of the genes[42] or for common functional content.[43]

The major result of the explorative analysis is essentially a list of potential marker genes relevant for the disease or treatment under analysis. Since microarray data is error-prone, this list contains a lot of false positives. Thus, further filtering steps are commonly included in the analysis. Recent methods therefore aim at the correlation of the gene expression profiles with complementing sources of data such as pathway annotation, gene ontology (GO) categories, sequence analysis, clinical data, etc.[44-46]

Genes do not act as individual units; they collaborate in overlapping pathways, the deregulation of which is a hallmark for the disease under study. New bioinformatics tools have been developed that judge gene expression changes in the context of such *pathway analysis*. We have developed a method that judges the alteration of entire pathways with respect to two experimental conditions. This has been applied recently for the identification of pathways altered upon differentiation of inner cell mass and trophectoderm in the human blastocyst[47] and upon hormone-induced aging of the human skin.[48] The procedure is based on pathway annotation (for example provided by the Kyoto Encyclopedia of Genes and Genomes [KEGG] pathway database).[49] This information is then translated into a two-dimensional statistical test problem that involves Wilcoxon's signed rank sum test in order to compute a Z-score for each pathway that quantifies the degree of alteration across the different experimental conditions. The results of the pathway analysis in the latter study, for example, implicate the involvement of several metabolic pathways in the aging process, such as C21-steroid hormone metabolism, phospholipid degradation, prostaglandin and leukotriene metabolism, 2,4-dichlorobenzoate degradation, and fatty acid biosynthesis. Interestingly, pathways operative in neurodegenerative disease such as Huntington's disease (HD),[50,51] dentatorubral-pallidoluysian atrophy (DRPLA),[52] and amyotrophic lateral sclerosis (ALS)[53] also showed significant age-dependent expression changes.

Databases, standardization initiatives, and common platforms

It has been recognized that there is a fundamental need worldwide to share microarray data in order to correlate researchers' results with already published data. Since for such a task it is necessary to provide the raw data, large microarray databases have been set up as public repositories (for example the gene expression omnibus (GEO) from NCBI[54] and ArrayExpress from EBI[55]).

Functional annotation is provided by the GO consortium.[56] The aim of GO is to maintain a consistent, species-independent, functional description of gene products. GO terms have a defined parent-child relationship and form a directed acyclic graph (DAG). At the root of the GO are the three top-level categories—molecular function, biological process, and cellular component—which contain many levels of child nodes (GO terms) that describe a gene product with increasing specificity. There are several tools for mining these annotations. We have developed the GOblet server that computes GO-term graph annotation for DNA sequences comprising several different sequence databases.[57,58]

A particular data repository for neuroscience applications is the National Brain Databank, a publicly accessible gene expression repository for the collection and dissemination of results from postmortem studies of neurological and psychiatric disorders. The project has been developed by the Harvard Brain Tissue Resource Center (HBTRC) in collaboration with Akaza Research, as an online resource for the neuroscience community.

A further useful database for drug discovery and drug response screening is PharmGKB.[59,60] This database is a central repository for genetic, genomic, molecular, and cellular phenotype data and clinical information about people who have participated in pharmacogenomics research studies. The data includes, but is not limited to,

clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathway, metabolic, and transporter domains. Currently, information on 385 drugs and 22 different pathways can be reviewed.

Standardization and the development of standard operating procedures (SOPs), both for data generation and data analysis, are major issues in community initiatives. Whereas SOPs are widespread in experimental procedures, they are not available for the data-analysis part. Publications often report data analysis methods that are hard to reproduce. Thus, it has been worthwhile to develop some common analysis platforms. Besides commercial programs there have been powerful open-source platforms such as R/Bioconductor. These platforms contain standard statistical procedures, visualization methods, and methods for importing and exporting data. In a script-based language data analysis methods can be reported and easily reproduced.

The "druggable genome"

The detection of genes (or compounds) that have a particular molecular feature that makes them useful for measuring disease progression or effects of treatments can be enhanced by microarray analysis, provided there is a robust data analysis and correlation of the experimental outcome. Other functional genomics technologies are needed to complement the results obtained from microarrays. These technologies (such as proteomics, metabonomics, etc.) are inherent in standard drug screening workflows in pharmaceutical companies.[61] However, the flood of data produced is not easily handlable, and requires a new generation of computational tools that are more effective in managing the data and are able to embed the obtained result in a functional context.[62,63] Current treatments for most neurological disorders are either ineffective or minimally effective or produce severe side effects (for example antipsychotic medication in schizophrenia[64,65]). Thus, there is a clear need for better methods. A potential direction could be the development of compounds that effectively address the disease pathways driven by disease-related pathway identification methods.

It has been reported that the number of potential drug targets is fairly limited. An assessment of the number of genes that might serve as potential targets for drugs has been given recently.[66,67] Starting from the fact that there are well-known properties that define good drugs, the number of potential drug targets is predictable. These properties can be summarized as[68]:

- presence of more than five hydrogen-bond donors
- molecular mass >500 d
- high lipophilicity
- more than 10 nitrogen and oxygen atoms.

These properties increase the likelihood of oral bioavailability of a compound, ie, what makes it a commercially viable drug. Looking at the binding sites on human protein sequences for such compounds, only approximately 400 potential targets have been identified. Extending these targets to all members of their relevant gene families, approximately 3000 molecular targets can be identified. Most of these genes belong to a few gene families such as G protein coupled receptors (GPCRs), serine/threonine and tyrosine protein kinases, and nuclear hormone receptors. The implications of these estimations are that the limited number of druggable targets will be well explored within the next decade, with chemical leads being available for most of them. Thus, there will be a shift from the development of leads to the investigation of the molecular consequences of the drug treatment in the individual patient.

Challenges in neuroscience applications

Drug discovery and treatment in neuroscience face specific challenges, in particular regarding the availability of tissue, poor diagnosis, complexity of brain tissue, and the lack of good model systems for drug target validation.[69] Tissue samples in neuroscience applications are mostly post-mortem brain samples from affected individuals. These samples typically reflect the end stage of the disease, which highly biases the material and makes it impossible to study early disease stages.[70] Furthermore, the patients have typically undergone some disease treatment, which has an influence on the gene expression. Thus, separating the effects of these treatments from the effects of the disease is extremely difficult. Here, animal models and tissue culture systems can help to identify marker genes and pathways for the disease, as is common in other studies. For example, in a recent work we have utilized a mouse model (Ts65DN[71]) for trisomy 21 in order to identify genes that show dosage imbalances with respect to aneuploidy.[29] Results for many genes (such as *APP*) could be extrapolated to human tissue samples. Good animal models allow the extraction of untreated brain material as well as material from control samples.

Rodent and (particularly) nonhuman primate models are primarily interesting in this respect.

Current research utilizes microarrays in several areas of neuroscience research, such as schizophrenia,[72,73] brain cancer,[74] Alzheimer's disease (AD),[75] and HD.[76] These studies compare gene expression changes in patient and control groups, and show that microarrays are valuable tools for the expression profiling of drug response in human individuals.

Interestingly, the latter study incorporated blood samples from patients and control subjects and revealed changes in blood mRNAs that reflect disease mechanisms observed in HD brain. Moreover, these alterations correlate with disease progression. For example, they were able to identify genes altered in blood from HD patients (such as *ANXA*, *CAPZA1*, *HIF1A*, *P2Y5*, *SF3B1*, *SP3*, and *TAF7*) that were also differentially expressed in HD postmortem brain. This work implies the potential of using easily accessible tissue such as blood for monitoring the progression of complex brain disorders.

Systems biology as a new research paradigm

Systems biology aims at the explanation of physiology and disease from the level of interacting components such as molecular pathways, regulatory networks, cells, organs, and ultimately the entire organism.[77] With the use of computer models for such processes *in silico* predictions can be generated on the state of the disease or the effect of the individual therapy. The new approaches are about to revolutionize our knowledge of disease mechanisms and of the interpretation of data from high-throughput technologies.[1]

These approaches are necessary, considering the increasing complexity of research. Often, several laboratories are working with different techniques on the same problem. A fundamental challenge is thus to search through the exhaustive set of data and extract meaningful information. Here, *in silico* experiments can be the basis for a more successful drug screening.

Furthermore, there is a fundamental need for integration rules and methods. Multiple databases exist, a variety of experimental techniques have produced gene and proteome expression data from various tissues and samples, and important disease-relevant pathways have been investigated. Information on promoter regions and transcription factors is available for many genes as well as sequence information. This information—although extremely helpful—cannot be utilized in a sufficient way because of the lack of integrative analysis tools. A fundamental aim of systems biology is the understanding of the underlying biological processes on the basis of this data.

Crucial for the step from qualitative, explorative data analysis to quantitative, predictive analysis is the combination of experimental data with the knowledge of the underlying biological reaction system. This approach makes it possible to come up with conclusions about the properties of the system, even those that are not subject to experiments or are not even amenable by any experimental approach. For this purpose we have developed the modeling and simulation system PyBioS.[78] With this system it is possible to construct models that are based on the topology of a cellular reaction network and adequate reaction kinetics. Based on this information the system can automatically construct a mathematical model of differential equations that can be used for subsequent simulation of the temporal behavior and model analysis. Particularly, information on the topology of biological systems is available from several databases (eg, KEGG). PyBioS provides interfaces to these databases that can be used for the construction of appropriate model prototypes. Models include metabolic pathways, signal transduction pathways, transport processes, gene regulatory networks, among others, and can be accessed via a Web interface.

Mathematical models for disease pathways have been developed, predominantly for cancer. Examples are general emergence of properties of signaling pathways[79] such as extended signal duration, threshold behaviors, etc, endodermal growth factor receptor (EGFR) signaling,[80-82] and the TNF alpha-mediated NF-kappa B-signaling pathway (NFκB).[83,84] Specific pathway models for neuroscience applications are currently rare. Nevertheless, an understanding of the dynamics of these diseases could help to develop strategies to halt them at the stage they have reached at detection, or to prevent them entirely.[85]

Conclusion

Despite the great uncertainties inherent in functional genomics techniques, they will be indispensable for future work in drug development and therapy monitoring. However, these techniques must be accompanied by solid support from data analysis. Bioinformatics, and to an increasing degree, systems biology, have key roles in

this process. The information that we can gain about a biological system (for example a disease process) appears in practice as an experimental observation, and research is restricted to the targeted molecular level and the precision of the experimental techniques in use. It is very likely that the range of this experimental granularity will increase in the coming years, utilizing heterogeneous techniques that target a biological question of interest at different points so that data integration becomes a major challenge for future biomedical research.

In the case of complex disease conditions it is clear that such integrated approaches are required in order to link clinical, genetic, behavioral, and environmental data with diverse types of molecular phenotype information and to identify correlative associations. Such correlations, if found, are the key to identifying biomarkers and processes that are either causative or indicative of the disease.

In order to screen the success of drug treatment in the individual patient, new generations of tools and research methods will be developed. These tools will enable us to perform the crucial step from qualitative to quantitative analysis. Systems biology is pointing in this direction. With its close connection of experimental data generation, predictive data modeling, and subsequent validation it holds the promise of providing computational tools capable of personalized treatment and therapy monitoring in the individual patient. ❑

The authors wish to thank Christoph Wierling for proofreading the manuscript and Sylvia Krobitsch for providing neuroscience literature.

REFERENCES

1. Hood L, Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. *Nat Biotechnol*. 2004;22:1215-1217.
2. Booth B, Zemmel R. Prospects for productivity. *Nat Rev Drug Discov*. 2004;3:451-456.
3. Weinshilboum R. Inheritance and drug response. *N Engl J Med*. 2003;348:529-537.
4. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531-537.
5. Gerhold DL, Jensen RV, Gullans SR. Better therapeutics through microarrays. *Nat Genet*. 2002;32:547-552.
6. Adler AS, Lin M, Horlings H, Nuyten DS, van de Vijver MJ, Chang HY. Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet*. 2006;38:421-430.
7. Mischel PS, Cloughesy TF, Nelson SF. DNA-microarray analysis of brain cancer: Molecular classification for therapy. *Nat Rev Neurosci*. 2004;5:782-792.
8. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. *Nat Genet*. 2005;37:S38-S45.
9. Lockhart DJ, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*. 1996;14:1675-1680.
10. Hughes T, Mao M, Jones A, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol*. 2001;19:342-347.
11. Gunderson KL, Kruglyak S, Graige MS, et al. Decoding randomly ordered DNA arrays. *Genome Res*. 2004;14:870-877.
12. Kuhn K, Baker SC, Chudin E, et al. A novel high-performance random array platform for quantitative gene expression profiling. *Genome Res*. 2004;14:2347-2356.
13. Hardiman G. Microarray platforms – comparisons and contrasts. *Pharmacogenomics*. 2004;5:487-502.
14. Lehrach H, Drmanac R, Hoheisel J, et al. Hybridization fingerprinting in genome mapping and sequencing. In: Davis KE, Tilghman S, eds. *Genome Analysis: Genetic and Physical Mapping.* Cold Spring Harbor, NY; 1990:39-81.
15. Lennon G, Lehrach H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet*. 1991;7:314-317.
16. Schena M, Shalon D, Davis R, Brown P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270:467-470.
17. Adjaye J, Herwig R, Herrmann D, et al. Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays. *BMC Genomics*. 2004;5:83.
18. Newton SS, Bennett A, Duman RS. Production of custom microarrays for neuroscience research. *Methods*. 2005;37:238-246.
19. Parrish ML, Wei N, Duenwald S, et al. A microarray platform comparison for neuroscience applications. *J Neurosci Meth.* 2004;132:57-68.
20. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*. 2002;18:405-412.
21. Tan PK, Downey TJ, Spitznagel EL, et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*. 2003;31:5676-5684.
22. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res*. 2005;33:5914-5923.
23. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33:e175.
24. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. *Genome Res*. 2002;12:325-332.
25. Wierling CK, Steinfath M, Elge T, et al. Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics*. 2002;3:29.
26. Steinfath M, Wruck W, Seidel H, Lehrach H, Radelof U, O'Brien J. Automated image analysis for array hybridization experiments. *Bioinformatics*. 2001;17:634-641.
27. Holleman A, Cheok MH, denBoer ML, et al. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *N Engl J Med*. 2004;351:533-542.
28. Quakenbush, J. Microarray data normalization and transformation. *Nat Genet*. 2002;496-501.
29. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*. 1979;74:829-836.
30. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc*. 1983:83:596-610.
31. Yang H, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variations. *Nucleic Acids Res*. 2002;30:e15.
32. Li C, Wong, WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci U S A.* 2001:98:31-36.

El perfil de expresión de la respuesta a los fármacos: desde los genes a las vías involucradas

La comprensión de la respuesta individual a un fármaco –que determina su eficacia y tolerabilidad en el organismo– es el principal cuello de botella en el desarrollo actual de fármacos y ensayos clínicos. La respuesta intracelular y el metabolismo, donde participan por ejemplo las enzimas del citocromo P-450, puede aumentar o disminuir el efecto de diferentes fármacos, dependiendo de la variante genética. La tecnología de microarrays ofrece el potencial para mapear la composición genética del paciente individual. Sin embargo, como los experimentos no son tan precisos deben acompañarse de un análisis sólido y consistente de los datos. Además, la investigación reciente apunta a la combinación de datos con metodología proveniente de modelos matemáticos que permiten una asistencia orientada a problemas en el proceso de descubrimiento de fármacos. Este artículo revisará el estado actual del conocimiento acerca de las plataformas de tecnología de arrays de ADN y los elementos básicos del análisis de los datos y la investigación bioinformática en el descubrimiento de fármacos. Para incrementar el análisis de un gen único, se presentará un nuevo método para la interpretación de los cambios en la expresión de los genes teniendo en cuenta todas las vías involucradas. Además se introducirá el concepto de biología de sistemas, como un nuevo paradigma para el desarrollo de fármacos, y se destacará nuestra reciente investigación acerca del desarrollo de un modelo y una plataforma de simulación para aplicaciones biomédicas. Finalmente se discutirán las potencialidades de la biología de sistemas para los modelos de respuesta a fármacos en el paciente individual.

Profilage de l'expression de la réponse au médicament : des gènes aux voies d'accès

La compréhension de la réponse individuelle au médicament, ce qui détermine son efficacité et sa tolérance chez le patient, est le principal goulet d'étranglement des essais cliniques et du développement des médicaments actuels. Le métabolisme et la réponse intracellulaires, par exemple à travers les enzymes du cytochrome P-450, peut soit augmenter soit diminuer l'effet des différents médicaments, selon la génétique. Des microéchantillons (microarrays) permettent de déterminer la configuration génétique de chaque patient. Ces techniques sont toutefois imprécises, justifiant une méthodologie précise et exigeante lors de l'analyse. De plus, la recherche récente permet un débit élevé de données avec des méthodes de modèlisation mathématique permettant de résoudre les problèmes ayant trait aux moyens de découverte des médicaments. Cet article concerne les techniques de pointe des plates-formes de technologie de microéchantillons d'ADN ainsi que les bases de l'analyse de données et de la recherche bio-informatique pour la découverte des médicaments. Nous présenterons une nouvelle méthode consistant à décrire les modifications de l'expression génétique au niveau de toute une cascade de réponses biologiques. Nous introduirons le concept de biologie des systèmes comme un nouveau paradigme pour le développement des médicaments et nous mettrons l'accent sur notre recherche récente, le développement d'une plate-forme de simulation et de modélisation pour les applications biomédicales. Nous discuterons du potentiel de la biologie des systèmes pour la modélisation de la réponse de chaque patient au médicament.

33. Irizarry RA, Bolstad BM, Collins F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2003;31:e15.

34. Draghici S. *Data Analysis Tools for DNA Microarrays*. Boca Raton, Fla: Chapman & Hall/CRC Press; 2003.

35. Kahlem P, Sultan M, Herwig R, et al. Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of Down syndrome. *Genome Res.* 2004;14:1258-1267.

36. Herwig R, Aanstad P, Clark M, Lehrach H. Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Res.* 2001;29:E117.

37. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95:14863-14868.

38. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A.* 1998;96:2907-2912.

39. Herwig R, Poustka AJ, Müller C, Lehrach H, O'Brien J. Large-scale clustering of cDNA fingerprinting data. *Genome Res*. 1999;9:1093-1105.

40. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. Paper presented at: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB); 2000; Menlo Park, California, USA.

41. Jain AK, Dubes RC. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall;1988.

42. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet*. 1999;22:281-285.

43. Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*. 2002;12:1574-1581
44. Gitton, Y, Dahmane, N, Baik, S, et al. A gene expression map of human chromosome 21 orthologues in the mouse. *Nature*. 2002; 420:586-590.
45. Rhodes DR, Barrette T, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*. 2002;62:4427-4433.
46. Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A.* 2004;101:9309–9314.
47. Adjaye J, Huntriss J, Herwig R, et al. Primary differentiation in the human blastocyst: Comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells*. 2005;23:1514-1525.
48. Makrantonaki E, Adjaye J, Herwig R, et al. Signalling and metabolic pathways associated with hormone-induced aging in human sebocyte cells in vitro. *Aging Cell*. 2006;5:331-344.
49. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resources for deciphering the genome. *Nucleic Acids Res*. 2004;32:D277-D280.
50. Luthi-Carter R, Strand AD, Peters NL, et al. Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. *Hum Mol Genet*. 2000;9:1259-1271.
51. Sipione S, Rigamonti D, Valenza M, et al. Early transcriptional profiles in huntingtin-inducible striatal cells by microarray analyses. *Hum Mol Genet*. 2002;11:1953-1965.
52. Luthi-Carter R, Hanson SA, Strand AD, et al. Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects. *Hum Mol Genet*. 2002;11:1927-1937.
53. Jiang YM, Yamamoto M, Kobayashi Y, et al. Gene expression profile of spinal motor neurons in sporadic amyotrophic lateral sclerosis. *Ann Neurol*. 2005;57:236-251.
54. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res*. 2005;33(Database Issue):D562-D566.
55. Parkinson H, Sarkans U, Shojatalab M, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*. 2005;33(Database Issue):D553-555.
56. Gene Ontology Consortium. The gene ontology (GO) project in 2006. *Nucleic Acids Res*. 2006;34(Database Issue):D322-D326.
57. Hennig S, Groth D, Lehrach H. Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Res*. 2003;31:3712-3715.
58. Groth D, Lehrach H, Hennig S. GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res*. 2004;32: W313-317.
59. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res*. 2002;30:163-165.
60. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenetics knowledge base. *Methods Mol Biol*. 2005;311:179-191.
61. Kramer R, Cohen D. Functional genomics to new drug targets. *Nat Rev Drug Discov*. 2004;3:965-972.
62. Kanehisa M, Bork P. Bioinformatics in the post-sequence era. *Nat Genet*. 2003;33:305-310.
63. Dobrin SE, Stephan DA. Integrating microarrays into disease-gene identification strategies. *Expert Rev Mol Diagn*. 2003;3:375-385.
64. Dunckley T, Coon KD, Stephan DA. Discovery and development of biomarkers of neurological disease. *Drug Discov Today*. 2005;10:326-334.
65. Evans WE, McLeod HL. Pharmacogenomics – drug disposition, drug targets, and side effects. *N Engl J Med*. 2003;348:538-549.
66. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Disc*. 2002;1:727-730.
67. Russ AP, Lampel S. The druggable genome: an update. *Drug Discov Today.* 2005;10:1607-1610.
68. Lipinski C, Lombardo F, Dominy B, Feeney P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. 1997;23:2-25.
69. Mirnics K, Pevsner J. Progress in the use of microarray technology to study the neurobiology of disease. *Nat Neurosci*. 2004;5:434-439.
70. Mirnics K, Middleton FA, Lewis DA, Levitt P. Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse. *Trends Neurosci*. 2001;24:479-486.
71. Reeves RH, Irving NG, Moran TH, et al. A mouse model for Down syndrome exhibits learning and behaviour deficits. *Nat Genet*. 1995;11:177-184.
72. Mirnics K, Middleton FA, Marquez A, Lewis DA, Levitt P. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron*. 2000;28:53-67.
73. Middleton FA, Mirnics K, Pierri JN, Lewis DA, Levitt P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J Neurosci*. 2002;22:2718-2729.
74. Mischel PS, Cloughesy TF, Nelson SF. DNA microarray analysis of brain cancer: molecular classification for therapy. *Nat Rev Neurosci*. 2004;5:782-792.
75. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor response. *Proc Natl Acad Sci U S A.* 2004;101:2173-2178.
76. Borovecki F, Lovrecic L, Zhou J, et al. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci U S A*. 2005;102:11023-11028.
77. Butcher E, Berg EL, Kunkel EJ. Systems biology in drug discovery. *Nat Biotechnol*. 2004;22:1253-1259.
78. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. *Systems Biology in Practice*. Weinheim, Germany: Wiley-VCH; 2005.
79. Bhalla US, Iyengar R. Emergent properties of networks of biological signaling pathways. *Science*. 1999;283:381-387.
80. Wiley HS, Shvartsman SY, Lauffenburger DA. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol*. 2003;13:43-50.
81. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol*. 2002;20:370-375.
82. Oda K, Matsuoka Y, Funahashi A, Kitano H. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Sys Biol*. 2005;1:2005.0010 Epub 2005 May 25.
83. Cho KH, Shin SY, Lee HW, Wolkenhauer O. Investigations into the analysis and modeling of the TNF alpha-mediated NF-kappa B-signaling pathway. *Genome Res*. 2003;13:2413-2422.
84. Hoffmann A, Levchenko A, Scott ML, Baltimore D. The I-kappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science*. 2005;298:1241-1245.
85. Aßmus HE, Herwig R, Cho KH, Wolkenhauer O. Understanding the dynamics of biological systems: roles in medical research. *Expert Rev Mol Diagn*. 2006. In press.