ELSEVIER

Achieving confidence in mechanism for drug discovery and development

Zach Pitluk and Iya Khalil

Gene Network Sciences, Inc., 10 Canal Park, Cambridge, MA 02141, United States

Decisions in drug development are made on the basis of determinations of cause and effect from experimental observations that span drug development phases. Despite advances in our powers of observation, the ability to determine compound mechanisms from large-scale multi-omic technologies continues to be a major bottleneck. This can only be overcome by utilizing computational learning methods that identify from compound data the circuits and connections between drug-affected molecular constituents and physiological observables. The marriage of multi-omics technologies with network inference approaches will provide missing insights needed to improve drug development success rates.

Introduction

The current business environment represents the 'best of times and worst of times' for drug discovery [1]. On the one hand, there is an exploding market demand from an aging population that has unparalleled unmet medical needs. On the other hand, there are strong demands for cost controls from payers, while the industry struggles to produce novel drugs and remain profitable.

Part of the research timeline dilemma is driven by pharmaceutical companies' dramatic improvements in reducing mortality and morbidity resulting from common diseases. Anti-hypertensive medicines, cholesterol reducing medicines, and strict glycemic control have saved millions of lives and billions of dollars, and provide a road map to future successes. Both hypercholesteremia and hypertension have excellent biomarkers and multiple points of therapeutic intervention. However, this information was built up painstakingly over decades, not within months or years. Documenting the complexity of the cholesterol pathway earned Brown and Goldstein a Nobel Prize nearly 25 years ago [2]. Furthermore, historic long-term prospective studies, such as the Framingham Heart Study [3], were pivotal in linking cholesterol to the pathology of myocardial infarction. The results of such studies often have limitations; the problem still remains that cholesterol will predict only one in two myocardial infarctions [4].

Corresponding author: Pitluk, Z. (zpitluk@gnsbiotech.com), Khalil, I. (iya@gnsbiotech.com)

In the current economic and scientific research environment, the timeline for companies to translate basic research into products is on the order of years, not decades. The economic challenge for drug discovery is to repeat the current success stories with two significant business drivers: a unique mechanism of action and first in class status [5]. Both of these objectives require the ability to understand the molecular mechanism by which a therapy causes the disease pathophysiology to be changed (mechanism of efficacy and toxicity, hereafter 'MoE-T', as defined below). Therefore, one of the fundamental issues in drug discovery and development is the question of how to move from the painstaking process of assembling and analyzing a mass of correlating evidence and data accumulated through the drug development process to a more nimble process of deriving causal relationships for discovery of compound MoE-T. MoE-T is the molecular mechanism by which a therapy causes (or fails to cause) the disease pathophysiology to change and the unintended side effects produced as a result of targeting the diseased state.

The difficulty in determining MoE-T is caused in part by a lack of understanding of on-target and off-target effects within the context of the complex cellular machinery. These effects can cause activation of unanticipated mechanisms or even interference with normal cellular function. Even low affinity events become likely, or even probable, given that total cellular protein levels are more than a thousand fold greater than most

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.10.001

targeted proteins [6]. Thus, understanding drug MoE-T at the molecular and physiological level necessitates being able to quantitatively predict the cause and effect relationship of cellular drug action. Determination of MoE-T is further confounded by the translation of results from one *in vitro* system to another, into the animal models that are used to select drug candidates and finally to human clinical trials. Understanding causal relationships between drug-affected biomolecules and pathophysiology in the cell, animal, and human systems used to advance a drug will go a long way in addressing these translational issues.

From correlations to causality

There are two crucial areas where causality is important in drug discovery process: target identification and validation, and translation into humans. Traditionally, targets were identified by studying biochemical changes in the diseased states. These biochemical changes were taken as direct causal evidence that the biochemical changes were driving the disease process. This approach yielded many successes with single-gene based diseases, for example, factor VIII for hemophilia [7], GBA for Gaucher's disease [8], BCR-ABL for chronic myeloid leukemia [9] and many others [10].

However, many diseases are multifactorial and not amenable to straightforward biochemical dissection. This highlights the first fundamental challenge for drug development: the discovery of biological target(s) for complex diseases with multiple causal mechanisms. Target identification and validation has proven particularly challenging for diseases such as cancer, where translational models are inadequate, and has prompted major initiatives for discovering genetic causes of cancer directly from genomic analysis of human tumors [11]. An elegant approach to determining targets that drive the disease state is to take advantage of genetics to gain insights into complex diseases on the basis of studying the inheritance of the complex trait [12–15]. In the past, the use of linkage maps and even SNPs has produced results at a frustratingly slow pace. Experiments conducted by Schadt *et al*. have leveraged genotypic, molecular profiling, and phenotypic data in order to identify key causal drivers of disease [16]. The advantage to this type of approach is that it can identify the genetic drivers of disease, and thus the value of the discovered targets is inherently higher.

Ultimately the criteria for success are not just in the targets selected, but in the performance of the drug candidate as it enters the clinic. In selecting a chemical entity, additional challenges arise because of the difficulty of accurately predicting pathophysiology in humans. One of the great success stories in modeling is the application of PK/PD models to improve the success rate of Phase I clinical trials [17]. The missing layer in current PK/PD is the ability to directly add in molecular information that enables insights into how the MoE-T is attained and differentiation of pharmaceutical therapies for development and marketing purposes.

We currently have the ability to measure and observe a large number of cellular or metabolic entities that change in response to therapy using 'omics' technologies [18–20]. While 'omics' technologies do not provide complete coverage of the various bioactive molecules in a cell/tissue, they do provide enough coverage such that it should be possible to identify salient causal outcomes, particularly when these approaches are combined with model systems that capture disease relevant complexity [21]. From the perspective of developing medicines, the entire process of hypothesis generation needs to highlight the essential components that connect molecular impact of medicine to phenotypic response on a timescale of days to weeks. When interpreted with standard statistical methods and combined with pathway informatics approaches, 'omics' scale experiments still suggest too many possibilities to test efficiently and are restricted to interpretation within the context of pathways and models of pathways derived from known biology. Unfortunately, known pathways are constructed by aggregating data from many diverse biological contexts (organisms, cell types, and experimental conditions) and are limited to what has been reported in the literature; the causal relationships derived from them are not necessarily relevant to how the compound is working in the particular disease context of interest. Taken together, pathway databases provide few insights when a compound is affecting previously uncharacterized targets and/or causing new biological phenomena.

Ultimately, addressing the challenges of discovery of MoE-T necessitates methods for learning models directly from compound-generated data (a combination of molecular profiling and phenotypic/physiological response measurements). This enables the determination of mechanism directly from the model systems used within the decision points along the drug development pipeline from the initial *in vitro* cell lines screens to *in vivo* animal and human systems [22] (Figure 1).

Learning models from data

Models learned from data can encompass quantitative causal relationships between independent variables such as a compound of interest to internal mechanisms connecting dependent variables (e.g. measured DNA alterations, changes in mRNA, protein and metabolites) to phenotypic readouts of efficacy and toxicity. These relationships can be represented graphically with the independent, dependent, and/or phenotypic readouts that are nodes in the model and the interactions between them as directed edges in a network. In addition, models enable the prediction of quantitative effects via *in silico* simulations of the model. This is achieved by asking 'what if' questions of the model such as 'what happens if I administer the drug at a certain dose while knocking down a given gene', enabling identification of the key causal targets in the efficacy path of the compound, as opposed to cellular entities that are responding to the compound but do not affect the phenotype of interest.

Approaches for learning models from data fall into two classes: those that rely on deterministic reverse-engineering techniques and those that use a range of probabilistic learning methods [23–25]. Among these, inference algorithms based on Bayesian reverse-engineering approaches have shown promise in elucidating biological mechanisms in a variety of contexts [26–29]. The advantages of such methods include the ability to integrate heterogeneous datasets including omic and non-omic experiments across multiple levels of biological organization. In addition, they can be optimized to handle noisy biological data and can represent a variety of signaling events, including protein cascades, gene

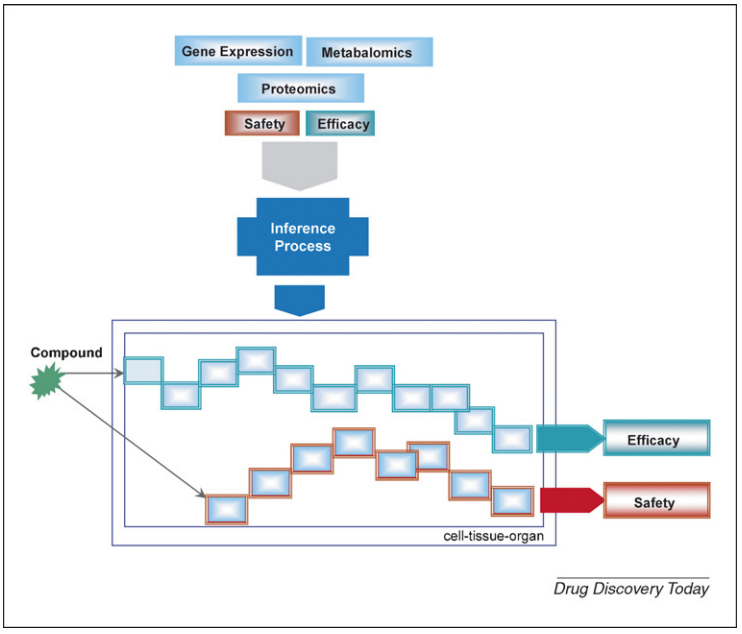

FIGURE 1

Inference process for analyzing profiling data. Models can be directly inferred from molecular profiling and phenotypic data via the inference process described in this review. Such models are able to predict new biological mechanisms and can encompass quantitative causal relationships between a compound of interest to measured DNA alterations, changes in mRNA, protein, and metabolites to phenotypic readouts of efficacy and toxicity.

expression networks, metabolic response, and combinations thereof.

Learning models from data starts by representing each of the measured attributes (e.g., the expression level of genes) in a given dataset as a random variable or probability distribution, whereby the actual measurement of the variable represents observed values under particular experimental conditions [30]. A model of the system within a probabilistic framework can be completely represented by a joint probability distribution over all random variables; this predicts the probability that the random variables can assume specified values (e.g. the probability that the expression of gene A and B is high while gene C is low). Fully solving such a joint probability distribution is intractable and requires a large number of parameters. Factoring this joint probability distribution into a product of local conditional probability distribution reduces the model to a product of terms, wherein each term has only a few parameters. In Bayesian network modeling, the model is also depicted graphically, whereby each node represents a measured random variable and directed edges (depicted as lines between the nodes) provide the conditional dependencies between these variables.

The task of learning models from data is then left to determining parameters for each of the local conditional probabilities or subgraphs and then determining the model (or models, as the data might be sparse or too noisy to converge on a single model) that best represents the data. The former is achieved by assigning a mathematical function to the subgraph and fitting its parameters to data. These functions can be, for example, discrete, linear, or non-linear, depending on the type of interaction and data being modeled (e.g., dose response drug experiments drive non-linear response in mRNA constituents measured via a microarray). The search over possible network topologies that could account for the data can be constructed by combining subgraphs until optimal combinations of subgraphs (networks) that best explain the process that gave rise to the data are determined. Exhaustively searching the space of all possible models is time-prohibitive and requires global optimization algorithms that minimize the cost of how well each candidate structure agrees with the observed data [31,32].

Several applications have taken advantage of the Bayesian framework for learning models from omic data. Sachs *et al.* sought to reconstruct the PKC protein signaling network using Bayesian network inference modeling from multi-phosphorylated protein flow cytometry data in primary human immune cells [33]. In this example, data from many thousands of cells under various stimulatory cues and inhibitory interventions were needed for learning the causal signaling events underlying the PKC pathway. Others have carried the approach to not only elucidate network topology structures, but also to understand the drug mechanism of action within the context of the inferred network topology [34]. Typically in drug development one is less interested in uncovering the detailed topology of the system and more interested in predicting key constituents involved in driving compound phenotypic response. It is not enough to determine that a set of molecular constituents changes because of the application of a compound; rather, the goal is to determine whether those markers cause a phenotypic outcome. This can be achieved via quantitative Monte Carlo simulations on the models that are reverse engineered from data. Woolf *et al.* were able to demonstrate, at least qualitatively, the effect of *in silico* inhibition of two nodes in a model reverse engineered from protein phosphorylation data was able to qualitatively demonstrate the impact of the perturbation on stem cell differentiation in mice [42].

Network inference framework for drug discovery and development

We have found that utilizing a combined reverse engineering and forward simulation framework approach provides the most powerful means for tackling questions relevant to the drug development process (Figure 2). Using this combined approach, we are able to assign quantitative metrics to hypothesis generated from the models. In this framework, models can be reverse engineered from drug–dose response molecular profiling and phenotypic data. Rather than inferring a single best fit model, we infer a distribution of models that are most likely given the data. This allows us to assess the accuracy of our predictions and assign a quantitative, probabilistic metric for the likelihood of the prediction being true, given the observed data. The metric is computed by determining the degree of consistency between the predicted models that vary because of noise and incompleteness of the data. In a given run, millions of models are propagated in order to arrive at the best fit solutions, necessitating the use of supercomputers to parallelize the process and reduce the timescale of model generation from weeks to less than a day. The resulting inferred models, in addition to including interactions between genes and phenotypic end points, explicitly represent (one or more) compounds through non-linear interactions that capture compound-to-gene dose response relationships. The explicit representation in the model of compound, gene, and phenotypic nodes allows for interrogation of the model through high-throughput forward simulations.

Collecting the appropriate drug–dose response from an *in vitro* cell line or an animal system to apply this approach merely requires that enough drug–dose treatments in the efficacy or toxicity range of the compound are represented in the training dataset. In addition, it is important that profiling and phenotypic measurements are collected under similar conditions, so as to be able to infer causal connections between them. For example, in order to determine the mechanism of action underlying a panel of kinase inhibitors using the above described approach, we applied three different doses of each compound roughly corresponding to IC_{10}, IC_{50}, and IC_{90} to various cell lines. Then, we measured global changes in transcript levels via Affymetrix arrays and in parallel observed subsequent changes in cell phenotype (the measurements are off set by different time points under the exact same treatments). We were then able to infer models that predicted causal gene markers impacting cell phenotype for each compound. Robust predictions can be gleaned from a dataset with as little as 40 treatments per cell line, making this approach economically feasible given the gain in insights it provides.

The framework described above can also be applied to molecular profiling and physiological data collected in clinical trials to determine compound mechanism of action and biomarkers in humans [35]. Here, the challenge is in procuring clinical samples of compound response from patients and establishing protocols for molecular profiling assays that limit the noise inherent in clinical samples. For diseases such as cancer, where molecular profiling tends to be done in the cancer tissue, additional challenges will stem from the inability to harvest enough sample mass, lack of biological replicates, and lack of an adequate normal control to the cancer phenotype. Nonetheless, researchers are applying Bayesian inference methods to increase the accuracy of classifying cancer patients by inferring the network structures

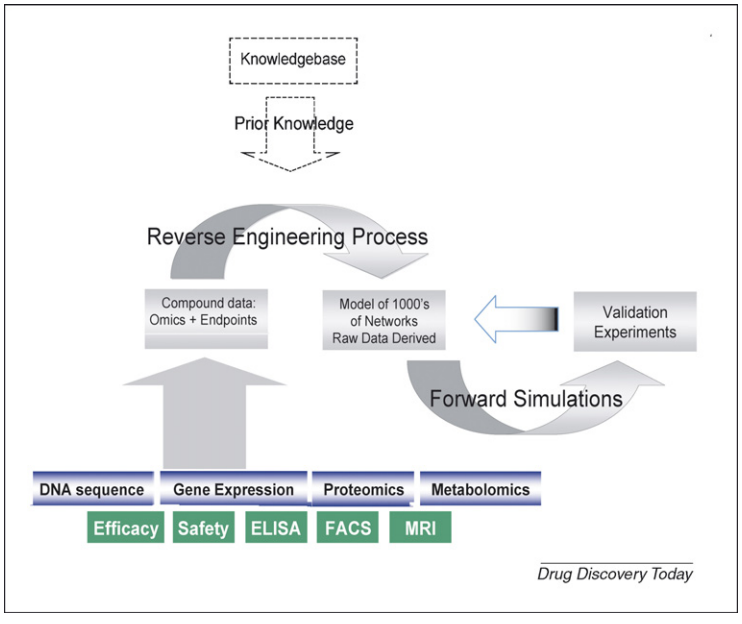

FIGURE 2

Reverse engineering and forward simulation process for discovering drug mechanism of action. In this process the first step consists of reverse engineering an ensemble of models (rather than a best fit single model) that provides the framework for generating quantitative causal predictions. Models can be generated *ab initio* or can be informed via prior knowledge from known biological interactions to help guide or improve the search for the optimal set of models that best explains the data. Once the models are generated, forward simulation technologies are used to interrogate the models in an automated high-throughput manner and test millions of possible hypotheses until a robust set of causal quantitative predictions are derived. These can be experimentally verified and used to further refine model predictions. In addition, predictions from the model can be analyzed in the context of known pathways to elucidate their relevance to drug mechanism of action.

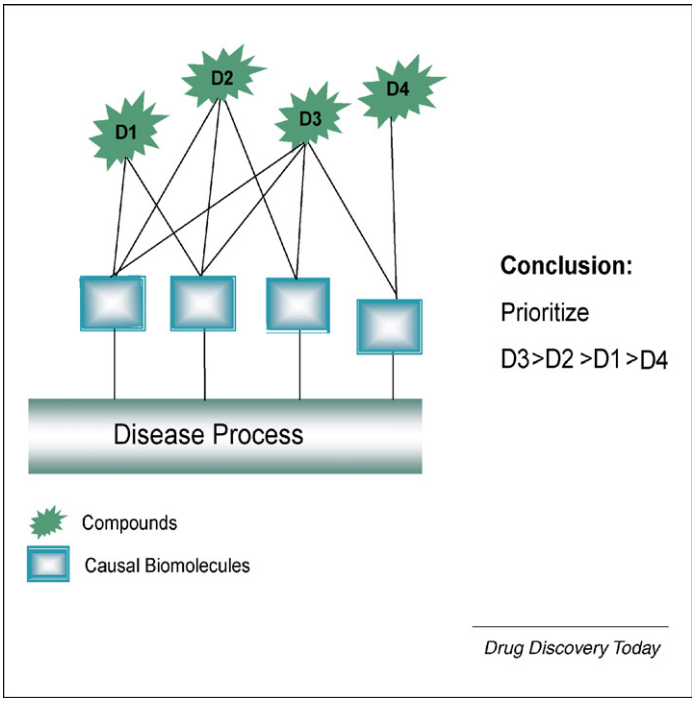

FIGURE 3

Network inference analysis of compound effects. Conceptual example of how network inference will be able to provide an objective view of how compounds work relative to each other within the context of disease. This insight will be highly useful for marketing decisions with regards to prioritizing compounds and which markets to enter.

that link molecular constituents to clinical covariate measurements such as response to therapy or prognosis [36]. The application of inference technology within the clinic setting may be more tractable in diseases for which it is possible to gain insights from analysis of serum or urine such as diabetes and where multiple samples can be collected from the same patient [37].

A new era for portfolio planning for the pharmaceutical industry

At the process level, network inference (NI) can provide an objective view of a drug candidate's 'off-target' and 'on-target' effects on a process as they progress through the entire pre-clinical and clinical drug discovery process, by identifying the genetic and protein entities that are essential for a drug's effect (Figure 3). NI can be specifically targeted at the translational issues of efficacy and toxicity by enabling the identification and comparison across species of the key mediators of signaling in causing endpoints.

In transitioning from Phase 0 through a Phase IIa study, the collection of clinical endpoints and biomics data can be used to model and provide quantitative predictions of the relationships between target coverage and efficacy outcomes in Phase I. The economic benefit would be potentially to reduce the number of steps in Phase I and to provide a starting set of metrics for patient selection in Phase IIa. Building on a network inference model from Phase I, the data collected from the first time in patient studies would enable the testing and refinement of molecular markers of efficacy, safety, and how they relate to the clinical endpoints. The NI model would allow significantly more accurate patient selection in Phase IIb and a refinement of the dosing protocol to avoid adverse events, both of which are significant factors in a successful Phase IIb study. Sheiner discussed the need for this approach a decade ago, describing the difference between a confirming and a learning trial [38]. The size of the learning trial would probably be 25–35 patients [based on our own unpublished modeling results]; this is within the current size of many Phase I trials. Additional thought will have to be given to appropriately design the study such that the appropriate variables are measured and sufficient replicate samples per patient are taken.

The second area of greatest economic impact is the application of the technology to studies after a drug is approved. Since only 3 in 10 drugs recover their initial investment [39], the pharmaceutical industry is under enormous pressure to find the best patient profiles to document the efficacy and benefits to patients. Using NI modeling, patient selection and understanding and the identification of issues arising from polypharmacy can be improved. Improved clinical outcome studies will help ensure that the right patients receive the right medicines with the best outcome for them. In a cost benefit analysis, many treatments already return a fourfold savings in terms of cost of treatment relative to avoided care costs [40]. The customer value proposition could be strengthened even further with the application of appropriate NI modeling studies, whereby data would include days of hospitalization, productivity measures for employers, and patient survey information. Using the best learning algorithms would allow pharmaceutical companies to deliver higher value to a burgeoning patient population.

Conclusions

There are four main pain points that network inference can address in drug discovery and development: (1) handing translational issues with inherent physiological differences between species, (2) providing an integrated framework for understanding diverse data types with respect to a process bounded by multiple physiological

endpoints in all phases of development, (3) deriving quantitative metrics that capture process accuracy in identifying off-target and on-target effects, and (4) adding value to the business management level by providing quantitative metrics to balance across and between therapeutic areas relative to competitor compounds.

The coming years will see a continued shifting of resources in the pharmaceutical industry toward collecting greater amounts of human data, since it is the main gap in resolving the translational medicine questions of efficacy and safety. This should result in fewer, but more accurate, experiments being conducted on faster timelines. A successful implementation of any type of modeling capability will require a different approach to experimental designs. Issues such as improving the powering of experiments will enable researchers to dig deeper into their data and identify causal relationships.

Currently, omics scale experiments are limited to an understanding at the correlation level, but would provide much more valuable insights if causal relationships could be extracted from the raw data. This will require that the platform be able to provide results on the timescale of days and weeks. A key innovation will be the use of forward simulations to extract working hypotheses in an automated fashion. The approach described above will allow an engineering approach to be applied to processes that are currently developed with little thought as to how to fully exploit the data for guidance in future experimental design beyond the inclusion of positive and negative controls.

In the realm of understanding the impact of combination therapies, the identification of key mediators is critical to ensure broad coverage across diseases, and to reveal the convergences in mechanism shared by different target classes. Taken together, the ability to make causal assessments of molecular information can transform every level of the development of medicines. The old, *ad hoc* methods currently in use cannot scale in volume with the flexibility and consistency that the new pharmaceutical business environment requires.

If network inference methods can deliver on the vision of selectivity, there will be a paradigm shift from the search for a master gene [41] to understand malfunctioning circuits. At its boldest, reverse engineering of models from GeneChips, metabonomic, and proteomic data may enable health sciences to explore the crucial intersection of nutrition and medicines. Personalized medicines may be made on demand at the pharmacist from a collection of already approved drugs and supplements. In the future, the practice of medicine will probably emphasize the individual, taking advantage of a deeper understanding of cause and effect.

References

1 Dickens, C. and McLenan, J. (1859) *A Tale of Two Cities*. T.B. Peterson and Brothers
2 Bilheimer, D.W. and Goldstein, J.L. *et al.* (1975) Reduction in cholesterol and low density lipoprotein synthesis after portacaval shunt surgery in a patient with homozygous familial hypercholesterolemia. *J. Clin. Invest.* 56, 1420–1430
3 Dawber, T.R. and Moore, F.E. *et al.* (1957) Coronary heart disease in the Framingham study. *Am. J. Public Health Nations Health* 47 (4 Part 2), 4–24
4 Stampfer, M.J. and Sacks, F.M. *et al.* (1991) A prospective study of cholesterol, apolipoproteins, and the risk of myocardial infarction. *N. Engl. J. Med.* 325, 373–381
5 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715
6 Mourant, J.R. and Yamada, Y.R. *et al.* (2003) FTIR spectroscopy demonstrates biochemical differences in mammalian cell cultures at different growth stages. *Biophys. J.* 85, 1938–1947
7 Antonarakis, S.E. and Youssoufian, H. *et al.* (1987) Molecular genetics of hemophilia A in man (factor VIII deficiency). *Mol. Biol. Med.* 4, 81–94
8 Barton, N.W. and Furbish, F.S. *et al.* (1990) Therapeutic response to intravenous infusions of glucocerebrosidase in a patient with Gaucher disease. *Proc. Natl. Acad. Sci. U. S. A.* 87, 1913–1916
9 Sawyers, C.L. (1999) Chronic myeloid leukaemia. *N. Engl. J. Med.* 340, 1330–1340
10 Pulst, S.M. (2003) Neurogenetics: single gene disorders. *J. Neurol. Neurosurg. Psychiatry* 74, 1608–1614
11 Varmus, H. (2006) The new era in cancer research. *Science* 312, 1162–1165
12 Brem, R.B. and Yvert, G. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755
13 Schadt, E.E. and Monks, S.A. *et al.* (2003) A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem. Soc. Trans.* 31, 437–443
14 Schadt, E.E. (2005) Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr. Opin. Biotechnol.* 16, 647–654
15 Rockman, M.V. and Kruglyak, L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.* 7, 862–872
16 Schadt, E.E. and Lamb, J. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37, 710–717
17 Ahmad, A.M. (2007) Recent advances in pharmacokinetic modeling. *Biopharm. Drug Dispos.* 28, 135–143
18 Ong, S.E. and Mann, M. (2005) Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* 1, 252–262
19 Stoughton, R.B. and Friend, S.H. (2005) How molecular profiling could revolutionize drug discovery. *Nat. Rev. Drug Discov.* 4, 345–350
20 Fan, J.B. and Chee, M.S. *et al.* (2006) Highly parallel genomic assays. *Nat. Rev. Genet.* 7, 632–644
21 Butcher, E.C. (2005) Can cell systems biology rescue drug discovery? *Nat. Rev. Drug Discov.* 4, 461–467
22 Aksenov, S.V. and Church, B. *et al.* (2005) An integrated approach for inference and mechanistic modeling for advancing drug development. *FEBS Lett.* 579, 1878–1883
23 Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science* 303, 799–805
24 Basso, K. and Margolin, A.A. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390
25 di Bernardo, D. and Thompson, M.J. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23, 377–383
26 Pe'er, D. and Regev, A. *et al.* (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 (Suppl 1), S215–S224
27 Smith, V.A. and Jarvis, E.D. *et al.* (2002) Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* 18, 216S–224S
28 Imoto, S. and Higuchi, T. *et al.* (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proc. IEEE Comput. Soc. Bioinform. Conf.* 2, 104–113
29 Imoto, S. and Kim, S. *et al.* (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.* 1, 231–252
30 Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press
31 Heckerman, D. (1999) A tutorial on learning with Bayesian networks. In *Learning in Graphical Models* (Jordan, M.I., ed.), MIT Press, Cambridge, Massachusetts, pp. 301–354
32 Mackay, D.J.C. (1999) Introduction to Monte Carlo methods. In *Learning in Graphical Models* (Jordan, M.I., ed.), MIT Press, pp. 175–204
33 Sachs, K. and Perez, O. *et al.* (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 523–529
34 Bose, R. and Molina, H. *et al.* (2006) Phosphoproteomic analysis of Her2/neu signaling and inhibition. *PNAS* 103, 9773–9778
35 Segal, E. and Friedman, N. *et al.* (2005) From signatures to models: understanding cancer using microarrays. *Nat. Genet.* 37 (Suppl), S38–S45

36 Gevaert, O. and Smet, F.D. *et al.* (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, e184–e190
37 Clayton, T.A. and Lindon, J.C. *et al.* (2006) Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature* 440, 1073–1077
38 Sheiner, L.B. (1997) Learning versus confirming in clinical drug development. *Clin. Pharmacol. Ther.* 61, 275–291
39 Grabowski, H. and Vernon, J. *et al.* (2002) Returns on research and development for 1990s new drug introductions. *Pharmacoeconomics* 20 (Suppl 3), 11–29
40 PHRMA (2006–2007). Annual Report
41 O'Malley, B.W. (2006) Molecular biology. Little molecules with big goals. *Science* 313, 1749–1750
42 Woolf, P.J. and Prudhomme, W. *et al.* (2005) Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics* 21 (6), 741–753