Hum Genet (2015) 134:479–495
DOI 10.1007/s00439-014-1501-x

REVIEW PAPER

An argument for mechanism-based statistical inference in cancer

Donald Geman · Michael Ochs · Nathan D. Price · Cristian Tomasetti · Laurent Younes

Received: 21 April 2014 / Accepted: 14 October 2014 / Published online: 9 November 2014

Abstract Cancer is perhaps the prototypical systems disease, and as such has been the focus of extensive study in quantitative systems biology. However, translating these programs into personalized clinical care remains elusive and incomplete. In this perspective, we argue that realizing this agenda—in particular, predicting disease phenotypes, progression and treatment response for individuals—requires going well beyond standard computational and bioinformatics tools and algorithms. It entails designing global mathematical models over network-scale configurations of genomic states and molecular concentrations, and learning the model parameters from limited available samples of high-dimensional and integrative omics data. As such, any plausible design should accommodate: biological mechanism, necessary for both feasible learning and interpretable decision making; stochasticity, to deal with uncertainty and observed variation at many scales; and a capacity for statistical inference at the patient level. This program, which requires a close, sustained collaboration between mathematicians and biologists, is illustrated in several contexts, including learning biomarkers, metabolism, cell signaling, network inference and tumorigenesis.

D. Geman (✉) · L. Younes
Department of Applied Mathematics and Statistics,
Johns Hopkins University, Baltimore, MD 21210, USA
e-mail: geman@jhu.edu

M. Ochs
Mathematics and Statistics, The College of New Jersey,
Ewing Township, USA

N. D. Price
Institute for Systems Biology, Seattle, USA

C. Tomasetti
Division of Biostatistics and Bioinformatics, and Department
of Biostatistics, Johns Hopkins University, Baltimore, USA

Introduction

The rationale for computational systems biology (Ideker et al. 2001) remains compelling: the traditional approach to biomedical research, experiments and analysis, done primarily molecule by molecule, is not suited to extracting system-level information at the scale needed to ultimately understand and model complex biological systems. Studying these systems in detail is now becoming possible due to data supplied by high-throughput technologies for genomics, transcriptomics, protemomics, metabolomics and so forth. Understanding the coordinated behavior and functional role of these many interacting components requires a principled and network-centered quantitative approach. In addition, "systems medicine" can reveal the perturbed structure of living systems in disease (Hood et al. 2004) as well as improved methods for disease diagnosis and treatment (Auffray et al. 2009; Hood et al. 2014).

This global view and quantitative research strategy has been widely adopted, and "computational" methods are now abundant in processing genomic signals, genome-wide association studies, inferring networks, discovering biomarkers, predicting disease phenotypes and analyzing disease progression. As promoted in Ideker et al. (2001), biomedical applications frequently involve "computer-based" models and simulation, and the development of bioinformatics tools and algorithms. Accordingly, survey articles about "translational bioinformatics" typically recount exemplary studies using

techniques from machine learning and statistics applied to specific subtasks (Altman 2012; Kreeger and Lauffenburger 2010; Butte 2008). Such techniques include new methods for stochastic simulation, mass action kinetics, data clustering, de-convolving signals, classification, testing multiple hypotheses, measuring associations, often borrowing powerful tools from computer science, biophysics, statistics, signal processing and information theory (Anderson et al. 2013).

Fully realizing the quantitative "systems" program in molecular medicine entails going beyond computer-based and bioinformatics tools. It requires designing mathematical and statistical models over global configurations of genomic states and molecular concentrations, and learning the parameters of these models from multi-scale data provided by omics platforms (Anderson et al. 2013; Auffray et al. 2009; Cohen 2004). Also, achieving a realistic balance between fidelity to fine-scale chemical dynamics and consistency with patient-level data necessarily requires a level of abstraction and generalization (Pe'er and Hacohen 2011).

Moreover, to have clinical relevance in complex diseases such as cancer, a mathematical model must provide for decision making at the individual patient level, including, for example, distinguishing among disease phenotypes, generating model-based hypotheses, and predicting risk and treatment outcomes (Altman 2012). Models can then be validated by the observed accuracy and reproducibility when ground truth is available, as well as more subjective factors such as the interpretability of the decision rules in biological terms. As a result, we argue here that most useful mathematical models for personalized molecular medicine, and cancer in particular, should accommodate at least three fundamental factors:

1. *Mechanism* The *causal* implications among biomolecules and phenotypes.
2. *Non-determinism* The inherent "stochasticity" in genetic variation, gene regulation, RNA and protein expression, cell signaling and disease progression.
3. *Inference* Generating predictions which are consistent with population statistics and identify individual disease phenotypes.

This paper is then largely a perspective on research strategy rather than a report of new results or even a review of existing ones. We argue for developing mechanism-based, statistical models and inferential procedures; similar arguments, more biologically oriented, are forcefully made in Pe'er and Hacohen (2011). "Statistical" is interpreted in a wide sense to accommodate statistical learning, whereby decision rules are induced from omics data using machine learning algorithms, and probabilistic modeling, for instance of the states of signaling molecules, the accumulation of mutations and tumor growth. Most existing statistical methods lack systematic hardwiring of biological mechanism which is necessary to improve accuracy and stability by limiting model complexity and to develop connections with existing biology. Conversely, few existing probabilistic models of networks or disease progression which do embed mechanism simultaneously allow for statistical inference. Recent exceptions include Vaske et al. (2010), Vandin et al. (2012), and naturally there are advantages to purely data-driven approaches when mechanistic information is lacking or scarce, for example, in generating initial insights and conjectures for rare cancers.

To illustrate these objectives, consider the case of network modeling. Understanding the role of specific genetic variants, transcripts and other gene products in health and disease requires identifying the main physical and causal interactions as a wiring diagram, sometimes referred to as "network topology". Yet no wiring diagram, no matter how richly annotated, is itself a "mathematical model", and a deep understanding also requires a global statistical characterization as well as an appreciation for network dynamics. Not all combinations of individual molecular states are equally likely; some configurations are observed far more often than others, and the favored states in health and disease are markedly different. A statistical model quantifies the likelihoods of molecular concentrations, not just individually but collectively as a multivariate probability distribution. This can be "translated" to practice by decision-making based on likelihood ratio tests, comparing the likelihoods of the observed data under various phenotype-specific probability models, or in a Bayesian framework by incorporating population statistics.

In summary, in our view, there is not enough global mathematical modeling in bioinformatics and computational systems biology, nor is there enough biology or statistics in existing mathematical representations. Getting mathematics, mechanism and inference simultaneously into the story requires persistent collaboration between mathematicians and biologists to select appropriate mathematical representations and inferential tools for a given medical context as well as identify the underlying context-specific biological mechanisms (Auffray et al. 2009; Butte 2008; Rejniak and Anderson 2012). Adding clinicians to these interdisciplinay teams can add tremendous value as well because focusing analyses on the pressing clinical questions is a major driver of eventual impact.

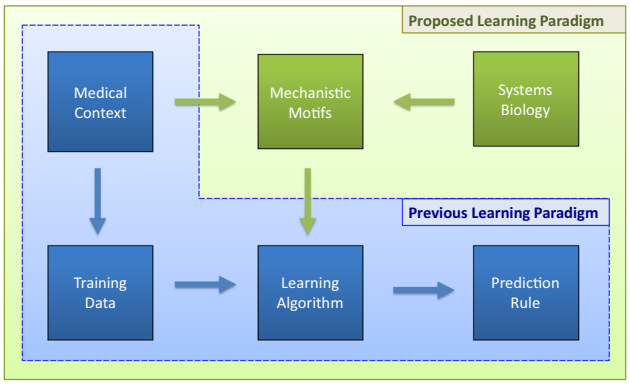


Fig. 1 The standard machine learning paradigm is depicted by the four *blue boxes*: depending on the classes or phenotypes under study ("medical context"), the input to a learning algorithm is training data consisting of samples from each class and the output is a prediction rule (classifier) for assigning a class to a new sample. Learning is then purely data-driven and often a "black box." The proposed modification adds the two *green boxes*: the learning algorithm restricts selection of the classifier to rules derived from context-dependent biological motifs; this constrains data-driven search by embedding mechanism and elucidates decision-making

Predicting disease phenotypes

For 15 years now, machine learning methods applied to omics datasets have yielded signatures and prediction rules that potentially discriminate among cellular and clinical phenotypes, facilitating enhanced detection and decoding disease processes, and prediction of clinical outcomes and response to therapy (Schadt and Björkegren 2012). Moreover, due to the considerable variability in the expression of individual genes or proteins among samples from the same phenotype, statistical learning (Hastie et al. 2009) is currently the core methodology for identifying predictors from high-throughput data. The standard procedure is illustrated in the blue boxes of Fig. 1: a prediction rule, which is a function that maps an observation vector (e.g., a gene expression profile) to one of the several classes (e.g., disease phenotypes), is learned or "induced" more or less directly from correctly labeled sample observations (e.g., a patient cohort) using a particular learning algorithm, often available as an R package which can be applied to data from any problem domain. Any biological analysis is usually post-learning, for instance, exploring associations between the features (e.g., genes) selected and the phenotypes.

For personalized medicine, the ultimate goal is to implement such procedures into assays to predict patient outcomes in the clinic. However, with the exception of a few FDA cleared assays for clinical use in cancer (Li et al. 2013; Cronin et al. 2007; Bender et al. 2009), the molecular-based predictors and signatures derived from statistical learning have largely not yet translated well to clinical use (Paik 2011; Marchionni et al. 2008a; Altman et al. 2011; Evans et al. 2011; Winslow et al. 2013), a situation that was recently evaluated by the U.S. Institute of Medicine (Omenn et al. 2012). Attributed reasons include insufficient accuracy, robustness and transparency; the difficulty of validating the "added value" beyond conventional clinical predictors (Boulesteix and Sauerbrei 2011); and perhaps a lack of incentive to engage in the complex and expensive process of obtaining clearance. These sobering observations suggest revisiting current strategies for learning with omics data.

The challenge of statistical learning in high dimensions

Many factors contribute to the limitations and under-performance of omics-based tests (Sung et al. 2012). Some concern inadequate study design (Simon 2006) and some concern data quality since high-throughput data are often strongly impacted by batch effects (Leek et al. 2010), reducing biomarker reproducibility. Moreover, significant biological variation is encountered from study to study for data collected for the same phenotype due to the underlying population heterogeneity. Although these issues are unavoidable, more stable and reproducible classification rates can be obtained by replacing ordinary randomized cross-validation by cross-study validation (Sung et al. 2012). In the case of human cancers, these challenges are being increasingly mitigated by large consortium efforts to catalog genomic states of human cancers, such as The Cancer Genome Atlas (TCGA) (Cancer-Genome-Atlas-Research-Network 2013).

In our view, the core challenge for translation-oriented statistical learning lies elsewhere, in two fundamental and related issues: instability and abstraction.

1. *Instability* The primary cause of the lack of reproducibility commonly observed with predictors learned from omics data is overfitting. This is manifested in practice by study-to-study differences in lists of discriminating biomarkers and highly variable accuracy on independent test data despite high reported accuracy in the samples used for discovery ("training"), contributing to the failure of clinical biomarkers (Simon et al. 2003; Kern 2012). The technical reasons for this instability can be analyzed mathematically and are attributed to the so-called *curse of dimensionality* and bias-variance dilemma (Geman et al. 1992), and the closely related *small n large d problem*. For omics data, the latter means that the number of samples n, e.g., expression profiles, available for learning predictors is often small relative to the number of potential biomarkers d, e.g., number of transcripts per profile. The most effective

way to enhance stability is to restrict the complexity of decision rules by hardwiring severe constraints into the discovery process.

2. *Abstraction* Most statistical learning algorithms are fundamentally data-driven rather than hypothesis-driven, having been developed in other domains and imported into computational biology. These learning algorithms, such as neural networks (Khan et al. 2001), random forests (Boulesteix et al. 2003), support vector machines (Yeang et al. 2001), boosting (Dettling and Buhlmann 2003), and linear discriminant analysis (Tibshirani et al. 2002) yield complex and abstract decision rules involving a great many components and non-linear relationships, and the search for discriminating structure is usually not informed by a priori domain knowledge (Varadan et al. 2012). Rather, biological context and interpretation only enter through post-hoc analyses of parameters and genes assigned in the decision rules. Consequently, these rules generally lack the mechanistic underpinnings necessary to carry meaning for biologists and clinicians, for example, to generate testable hypotheses or implicate therapeutic alternatives.

The "small *n*, large *d*" problem seems here to stay due to a variety of factors, including the prohibitive cost of dramatically increasing the number of patient profiles, patient stratification into smaller subgroups for personalized medicine, and the likely increase in *d* as measurement technologies improve and new classes of biomolecules are added to high-throughput experimental platforms. Reducing the number of candidate omics features by statistical filtering for phenotype associations can mitigate overfitting, but such methods have been of limited success (Porzelius et al. 2011). Statistical learning from even the largest datasets, like those used for Genome Wide Association Studies, can exhibit overfitting, especially when looking for combinations of rare variants in relation with phenotypes.

Here we argue that the absence of a clear-cut biological interpretation for the decision rules produced from using standard algorithms in statistical learning with omics data is a significant impediment to medical applications. Despite a large body of work, a solid link with potential mechanisms is notably missing, which seems to be a necessary condition for "translational medicine" (Winslow et al. 2012).

Prior biological knowledge

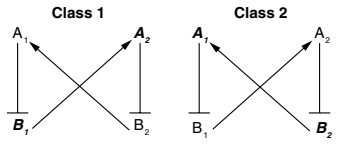
Instability and abstraction can be simultaneously addressed by reducing model complexity informed by a priori biological knowledge. Systematically leveraging prior information about biological networks will simultaneously severely constrain the search for predictive models to those with a potentially mechanistic justification and overcome the technical limitations inherent in *tabula rasa* statistical learning.

There have been recent efforts to move away from purely data-driven learning. Perhaps the most straightforward way is to restrict decision rules to signatures composed of genes previously annotated to the disease or "significantly differentially expressed" among the phenotypes of interest. However, such set-based techniques predominately restrict the use of biological knowledge to grouping information, frequently ignoring gene and protein neighborhood relations, and maintain the complexity of the decision rules. Other recent studies move closer towards mechanism by incorporating prior knowledge of molecular interactions in networks and cellular processes into the feature selection and prediction rules (Johannes et al. 2010; Zhu et al. 2009; Pan et al. 2010; Binder et al. 2009; Li and Li 2008) or identify differential expression at the level of pathways rather than individual genes (Khatri et al. 2012; Eddy et al. 2010; Subramanian et al. 2005). Selections are largely based upon curated gene sets and literature and these studies have reported improvements in cross-study validation (Lottaz and Spang 2005; Wei and Li 2007; Abraham et al. 2010; Chen et al. 2010). However, these networks are usually applied across phenotypes, regardless of the context in which they were learned. For these reasons it is not surprising that such networks sometimes provide only equivalent predictions to randomized networks, such as was observed for breast cancers (Staiger et al. 2012).

Embedding context-specific mechanism

We advocate hardwiring phenotype-dependent mechanisms specific to cancer pathogenesis and progression directly into the mathematical form of the decision rules. One strategy is to tie the decision rules to circuitry involving microRNAs (miR), transcription factors (TF) and their known targets that control key cellular processes in cancer (Mendell 2005; Hobert 2008; Croce 2009). Regulatory circuits of distinct topology include feed-forward loops and feedback loops, and one can attempt to identify such network motifs in signaling pathways and biochemical reactions intimately linked to the cancer phenotypes under study. For instance, for metastatic recurrence one could focus on TFs, miRs, and pathways involved in epithelial to mesenchymal transition and cell plasticity.

Basically, then, we are using "motif" in the sense of a small directed subnetwork of a generally much larger regulatory, signaling or metabolic network. Two points should be emphasized. First, a motif by itself does not determine



	On	Off
Switch 1	$A_1 < A_2$	$A_1 > A_2$
Switch 2	$B_1 > B_2$	$B_1 < B_2$

Fig. 2 Due to the depicted activation and suppression patterns, we might expect that A_1 is very likely to be expressed less than A_2 in class 1 and vice-versa in class 2. The comparison between the expression levels of B_1 and B_2 goes in the opposite direction: the event that B_1 is expressed less than B_2 might be far less likely in class 1 than in class 2. Hence, this motif generates two "switches", both likely to be "on" in class 1 and "off" in class 2

a decision rule for discrimination; doing so requires learning a mapping from the possible states of the motif, usually mutated genes or molecular concentrations, to the phenotypes of interest. The final decision rule may involve multiple motifs. Whereas learning this decision rule is data-driven, the set of possible signatures has been vastly reduced, which is the "hypothesis-driven" aspect. An example with a circuit involving two miRNAs and two mRNAs is given below. Second, ideally the set of candidate motifs for decision rules would be known a priori, before statistical learning. Of course such knowledge may not always be available, in which case a tabula-rasa, data-driven approach may be necessary to learn candidate motifs, hopefully involving different data and/or experimental verification to reduce overfitting.

One might combine this strategy with assembling predictors from elementary and parameter-free building blocks. In fact, studies have shown that simplicity does not necessarily limit performance (Dudoit et al. 2002) and that prediction rules based on fewer genes and parameters can be as sensitive and specific as more complex ones.

Consequently, these building blocks could be as simple as "biological switches" based on two-gene comparisons (Geman et al. 2004; Xu et al. 2005; Tan et al. 2005; Edelman et al. 2009). For example, in Price et al. (2007), a reversal between the two genes Prune2 and Obscurin was shown to be an accurate test for separating GIST and LMS, two morphologically similar cancers that require very different treatments. The decision rule is sufficiently elementary to support a possible biological explanation: both modulate RhoA activity (which controls many signaling events), a splice variant of Prune2 is reported to decrease RhoA activity when over-expressed and Obscurin contains a Rho-GEF binding domain which helps to activate RhoA. Extensions to aggregating multiple switches have been used to predict treatment response in breast cancer (Weichselbaum et al. 2008) and acute myeloid leukemia (Raponi et al. 2008), grade prostate cancers (Zhao et al. 2010), and prognosticate lung cancer (Patnaik et al. 2010). Nonetheless, these decision rules were learned from data using a largely unconstrained search of all possible switches, and hence do not illustrate an explicitly motif-driven discovery process.

Consider the bi-stable feedback loop depicted in Fig. 2. The two "classes" represent two phenotypes. Suppose, for example, that molecules A_1, A_2 (resp. B_1, B_2) are from the same species, for example, two miRNAs (resp., two mRNAs), and letters in boldface indicate an "expressed" state. Given both miRNA and mRNA data, the decision would be based on the number of "on" switches; see Fig. 2. Such motif-based predictors could then be aggregated into more global and powerful decision rules by arranging the corresponding motifs according to an overarching organizational framework recapitulating the "hallmarks of cancer" (Hanahan and Weinberg 2000; Hanahan 2011).

Another powerful means to embed context-specific mechanism into statistical learning is to leverage known biochemistry. Consider the example of cancer metabolism. Reprogramming energy metabolism is a fundamental and widespread characteristic of cancer cells (Hanahan 2011). To grow and metastasize, cancers must undergo a metabolic shift to enable these behaviors. This is not just a statistical correlation or generally observed pattern—if a cell does not alter its metabolism to accommodate it, enhanced growth cannot happen because it would violate basic physical laws such as mass and energy balance. Thus, we immediately have a strong mechanistic foundation to study cancers by studying omics data in the context of metabolic networks. Alterations in cancer metabolism are also involved in therapeutic response, as altered expression of detoxification metabolic pathways is implicated in chemotherapy resistance (Zhang et al. 2005). The use of metabolic networks to provide mechanistic context to inference from high-throughput data will be considered in more detail in the following section.

Metabolism

Metabolism represents one of the best characterized processes in biology, and we now have a genome-scale mechanistic reconstruction of the underlying biochemistry in humans (Thiele et al. 2013). Metabolic networks themselves naturally integrate across multiple omics domains, including genomics, proteomics, and metabolomics. Many

decades of careful experimentation have gone into building these comprehensive biochemical models, providing a foundation for computational and mathematical strategies that leverage this knowledge to better inform statistical models for personalized medicine.

Metabolism lends itself well to building mechanistic models that can serve as a basis on which to build the types of mechanism-driven statistical models for which we are arguing herein. One approach that has proven very useful for modeling microbes (Price et al. 2004) and more recently human systems (Shlomi et al. 2008) is known as constraint-based modeling. Briefly, this approach is a means to evaluate the range of possible states a biochemical network can have subject to governing constraints (e.g., steady-state mass balance) and available data (e.g., uptake/secretion rates, what metabolites are available in the microenvironment, etc.). These types of models have very few parameters, or are parameter-free given the network structure, and thus can be applied in scenarios where fully parameterized models are not possible (as is usually the case). The key then is to link such models with high-throughput data and statistical learning to drive forward personalized medicine grounded in biological mechanism.

As was mentioned above, there is now a genome-scale metabolic reconstruction for humans (Thiele et al. 2013) encompassing over 7,500 metabolic reactions in a unified framework. Leveraging the mechanistic information in the global human metabolic network reconstruction, it is then possible to use data-driven approaches that utilize omics data to contextualize the most likely tissue and cell-specific metabolic networks, for which initial versions have now been done for most tissues and many human cell types (Agren et al. 2012; Wang et al. 2012; Shlomi et al. 2008), and to use these as the basis for simulation of capabilities using constraint-based modeling. These genome-scale models of metabolic biochemistry also exist for a number of human pathogens (Jamshidi and Palsson 2007; Chavali et al. 2008) and other members of the human microbiome (Levy 2013), enabling context-driven statistical learning for host pathogen interactions based on similar methods (Bordbar et al. 2010).

Genome-scale metabolic network models have already been used to guide interpretation of high-throughput data successfully in a number of different contexts (Hanahan 2011; Milne et al. 2009; Oberhardt 2009). In cancer, these models have been used to evaluate the hypothesis that the Warburg effect, one of the hallmarks of cancer (Hanahan 2011), trades off efficiency of ATP production as a primary means to drive cell growth (Shlomi et al. 2011). Tumors exhibit heterogeneous metabolic profiles, as demonstrated by the differential uptake and secretion of metabolites such as glucose, glutamine, lactate and glycine (Barrett et al. 2006; Folger et al. 2011). This heterogeneity has been demonstrated in breast cancer, as ER-negative breast cancer cells are more dependent on the serine synthesis pathway than ER-positive breast cancer cells (Frezza et al. 2011). Building genome-scale metabolic models for cancer has been the subject of intensive study recently, and initial validation screens have shown their ability to predict essential genes across a number of cancer cell lines (Folger et al. 2011). Genome scale metabolic networks have also been successfully used to identify potential selective drug targets (Jerby 2012). One of the most successful demonstrations to date used a metabolic model of renal cancer to discover that a disruption of heme biosynthesis was synthetically lethal with the loss of the metabolic enzyme, fumarate hydratase. This identified synthetic lethal pair provided an ideal opportunity to design an approach to kill cancers in patients selectively with a targeted therapy, and indeed this calculated interaction was then experimentally demonstrated (Frezza et al. 2011), an important demonstration of the capability to design a targeted therapy from a model-driven approach.

In model organisms, combining gene regulatory and metabolic networks has proven to be a powerful means to integrate statistical and mechanistic networks (Chandrasekaran and Price 2010; Covert 2004; Price et al. 2007). Most recently, it was shown that conditioning putative gene regulatory associations on a framework of biochemical mechanism represented in metabolism could significantly enrich overlap with gold-standard gene regulatory interactions (Chandrasekaran and Price 2013). While such an approach has not yet been applied for human cancer, it represents a fascinating avenue for exploration to leverage decades of work in elucidating mechanistic understanding of cancer metabolism for the purpose of better uncovering metabolic regulation through mechanism-guided statistical inference.

We can also utilize metabolic networks to provide metabolic context for studying genomic variants. For example, it is valuable to constrain searches for multi-genetic drivers of cancer using selected combinations based on known biochemical mechanisms of interaction. It is of course true that biasing models towards what is already known will inevitably miss important targets, which can be identified via a complementary and iterative process of data-driven discovery and subsequent experimentation. Metabolic networks are particularly amenable to constraint-based mechanistic modeling approaches because the biochemical reactions and the genes responsible for catalyzing those reactions are well characterized. Thus, we can use a mechanistic biochemical framework for the analysis of selected genetic variants. In particular, constraint-based modeling can be used to predict variants that cause defects in energy metabolism or the production of important molecules of interest. Moreover, metabolic networks enable the so-called

forward calculation (i.e., based on mechanism and not reliant on statistical inference from training data) (Brenner 2010) that can link genotype with phenotype and make patient-specific risk predictions. These network-based strategies deliver mechanism-rooted networks that provide testable predictions of sets of genetic variants.

To make this more concrete, consider the simplest types of aberration that we can examine in this context: loss-of-function mutations in metabolic enzymes. By blocking flux through the corresponding reactions in the metabolic model, we can simulate the effect of these mutations on the entire network. Alternatively, we can perform sensitivity analysis on the catalyzed reactions to determine the effects of impairing any particular enzyme on the functioning of the network as a whole, and relate this to identified variants and their effects towards cancer metabolism. Importantly, reconstructed metabolic networks can be studied to define correlated sets of reactions, or co-sets. These co-sets represent groups of reactions that must function together in metabolic networks under the constraints of mass conservation, charge conservation, and thermodynamic considerations (Jamshidi and Palsson 2006). More precisely, co-sets represent reactions that have steady-state fluxes that are perfectly correlated. Co-sets are often non-obvious, as the reactions within a set may often not be adjacent on a network map. Notably, co-sets are precisely mathematically defined functional modules of a network and identify genes whose products are collectively required to achieve physiological states. As such, perturbations affecting any gene belonging to the same co-set would be expected to lead to similar functional consequences. This provides a basis for linking different mutations in genomes to common "buckets" to reduce dimensionality and then we can use the networks to rationally link up these buckets to drive combinations in the smaller search space with mechanistic links.

Signaling networks

Networks of signaling proteins in cancer

An example of the importance of networks in biological systems is the role of signaling in cancer. The discovery of key cell signaling proteins, such as p53 and RAS, and their interactions radically altered our understanding of how cancer cells overcome internal and external restraints on growth and metastasis (Hanahan and Weinberg 2000; Hanahan 2011). These proteins form pathways, on the order of six or seven, so that there are many potential points of deregulation (i.e. proteins), and, in any individual patient and tumor, a different protein in the pathway may be affected and driving pathway deregulation (Parsons et al. 2008; Li et al. 2013). In addition, while the early focus was on mutation of tumor suppressors and oncogenes, the activity of proteins in the pathway may be driven by promoter methylation, amplification, miRNA targeting, and other potential changes targeting the gene or mRNA.

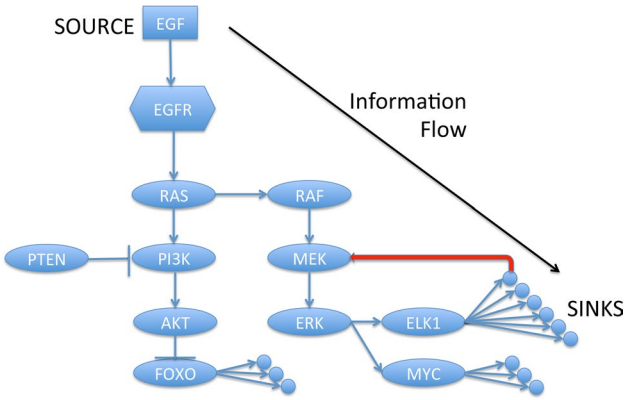


Fig. 3 A simplified model of a cell signaling process highlighting the flow from signals generated externally by epidermal growth factor (EGF) to the activation or repression of transcription. In addition, potential feedback in the form of expression of signaling repressor proteins is shown. Drivers that would make useful targets for intervention could lie anywhere within a pathway and be themselves the result of different molecular events (e.g., promoter methylation, mutation, gene amplification)

Given a goal of tailoring treatment to the individual tumor, we face a need to integrate diverse molecular measurements and interpret these in terms of pathway changes driving tumor growth and gene or protein aberrations that drive these pathways. We must then integrate gene-level molecular measurements to both identify aberrant pathway activity and deduce causality among the interactions among the proteins in the pathway.

Initial approaches relied heavily on expression data (technically transcription data as translation was not included), as microarrays provided the first widely obtained genome-wide measurements. Efforts focused on gene set analysis using the sets defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) or BioCarta databases (Kanehisa et al. 2002). But such analyses do not incorporate known biology, where gene expression is a downstream effect of cell signaling processes, which themselves are typically not driven by expression changes but by post-translational modification of low expression proteins. Therefore an appropriate causal model must instead ask how is expression driven by signaling and what is driving the observed expression changes.

Placing this within the context of a graphical model, we obtain something like Fig. 3. This is clearly a highly simplified view of signaling, and real networks will be significantly more complex, but it raises two immediate issues. Given even this simplified form, what can be learned based on measurements we can make and are there limits

to non-quantitative approaches, such as interpreting Fig. 3 visually as is typically done today?

The first issue on learnability has an immediate result given to us by epistasis. If all the information we have is downstream of a single protein that itself is downstream of another single protein, such as ERK and MEK, respectively, then activation of MEK by mutation cannot be distinguished from activation of ERK by mutation without additional data besides expression. Essentially, to identify drivers we will need to model the network mechanistically.

The second issue is much more complex. The fact that any measurements we actually make are inherently noisy both biologically and technically requires that the system must be viewed as containing a significant random component in each measurement, and therefore it is stochastic. This leads naturally to a result well known in medicine, that it is the overall systemic state of a patient that must be considered when viewing any individual laboratory value or test result.

An excellent example of the need to build a mathematical model is in the ability of signals to follow a parallel path when the primary path is blocked. In this case, though there is no "feedback" provided by protein interactions, the blocking of a signal in one of two branches downstream from a node leads to increased signal in the other branch through retroactivity (Wynn et al. 2011). Many biologists feel that when the expected response does not occur, that there is a component of the system that has not yet been discovered. While this is certainly possible, it cannot be stated coherently without a model of the existing state of knowledge capable of making quantitative predictions across multiple linked signaling pathways.

How then should one approach the issue of identifying drivers of aberrant signaling at the level of an individual tumor given the large number of different molecular measurements? One approach is to use a more realistic and cancer-type-specific graphical model similar to Fig. 3 as a prior for interpretation of the data. This substantially reduces the space of potential interactions and introduces a prior belief on the causal effects of molecular interactions (e.g., if RAS is active then RAF will be active). With this prior, inference takes the form of forming a tumor-specific posterior distribution that integrates the data relative to normal variation to infer points where upstream changes impact downstream readouts. For example, in one pathway methylation of PTEN could lead to loss of repression of FOXO, while elsewhere a mutation in RAF could activate MYC and ELK1.

Some work has begun to follow this integrated approach. The methods most limited in data integration incorporate interactome or curated pathway information into gene expression analysis (Liu et al. 2012a, b; Kim et al. 2011; Ochs et al. 2009). Other efforts have focused on identifying potential specific points of deregulation either by identifying deregulated subnetworks in the signaling pathways or using diverse molecular measurements to determine the specific potential drivers (Ulitsky et al. 2010; Ochs et al. 2014).

Overall, the most promising path to introduce mechanism into statistical models is through the capture of biological relationships within graphical models. For signaling, some progress has been made with the use of limited biological knowledge (Tuncbag et al. 2013; Ng et al. 2013; Wilson et al. 2013), but better collaborations between biologists and mathematicians are needed to adequately capture biology in the models.

Data-driven inference of network models

The general goal of elucidating the relationships among molecular species emerged quickly following the development high-throughput measurements (Eisen 1998; Butte 2000, 2003; Friedman 2004; Margolin et al. 2006). The analysis of correlation or mutual information between variables associated to gene expression data has led to multiple methods, like relevance networks, Gaussian graphical models and Bayesian networks to estimate an interaction graph among variables.

These methods are data-driven. They explore, at different levels of mathematical complexity, statistical relationships among variables. Basic approaches like relevance networks are limited to estimating graphs, placing an edge between variables that are considered to be directly related. Model-based methods pursue a more ambitious goal. They attempt to estimate a joint probability distribution among all the variables in the system that, within a class of statistical models, provides the closest approximation to the distribution of the observed data. The model class is generally associated to graphical models (Bayesian networks, Markov random fields), in which the pattern of conditional dependency among variables is represented by a directed or undirected graph, while the graph induces, in turn, a parametric representation of the distribution (Hartemink et al. 2005).

The task of learning both the graph and the associated parameters is referred to as *structure learning* in the graphical model literature (Neapolitan et al. 2004; Koller 2009). The difficulty of such an enterprise is, however, formidable. Disregarding computational challenges, which are serious, since the problem is NP-Complete, the parametric and combinatorial complexity of the underlying model class of graphical models makes any attempt at data-driven learning of network interactions with some reasonable accuracy simply impossible. Already with five or six variables, estimating networks based on typical sample sizes cannot be achieved without additional constraints on the

structure. Changing the data size by an order of magnitude would at best allow for the addition of a few more variables to the maximal size of networks that can be reliably estimated. One of the reasons for this is that there typically exist multiple network topologies, with similar complexity, that provide good approximations of the observable data. Even small variations in the data will make the optimal solution oscillate. This may not be a problem if the goal is limited to finding a good approximation of the joint probability distribution of the observed variables, but this is a serious impediment if one wants the observed structure to be mechanistically interpretable, allowing, for example, to predict the effect of network perturbations on the overall behavior of the system.

Indeed, one of the main appeals of probabilistic graphical models is that they can be used to analyze the effects of small perturbations on their overall behavior. For example, one may decide to knock out a variable (clamp it to 0) and measure the induced changes in the model. Here, we are not primarily interested in the statistical effect of clamping the variable, but on its mechanistic, or causal, impact, which, in general, cannot be inferred from population statistics. To take a simple example, imagine a system with two variables A and B such that A corresponds to a given mutation and B is associated to some viral disease, both variables taking values 0 or 1. Assume that $A = 1$ with probability p and that, conditionally to A, $B = 1$ with probability $(A + 1)q$ (so that the sensitivity to the disease is twice as likely when the mutation is present). Given that an individual has the disease ($B = 1$), the probability of mutation ($A = 1$) can be computed using Bayes rule and is equal to $2p/(p + 1)$. This comes from elementary statistical inference, and this rate can be estimated using samples of the population, simply dividing the number of occurrences of diseased mutants, divided by the total number of individuals with the disease. Now, imagine an experiment in which the disease is inoculated to the whole population, which corresponds to constraining $B = 1$ artificially. Then, the rate of individuals with mutation will not change, and remains equal to p . This mechanistically obvious statement cannot be inferred from statistical observations of the original population. In the absence of a mechanistic interpretation, one would have to actually perform the "experiment" (something referred to as an *intervention* in the causal inference literature) to be able to draw the conclusion.

More generally, a given stochastic phenomenon can be explained by a possibly large number of causal interpretations (Pearl 1988, 2000; Maathuis et al. 2009, 2010). Deciding between these interpretations must be based either on prior knowledge (Lee et al. 2002; Yoruk et al. 2011; Simcha et al. 2013) or on additional evidence (intervention) (Sachs et al. 2005). Since designing interventions, if even possible, can be extremely costly, the priority should be placed on the first option, that is, relying on as much biological expertise and evidence as possible in the design of a causal network, reducing the structure learning part to small perturbations, at most, of an initial hard-wired network.

Another issue that limits the usefulness of purely data-driven methods is the fact that statistical association does not necessarily correspond, even indirectly, to functional relationships. More precisely, while assuming that "molecular influences generate statistical relations in data" (Pe'er and Hacohen 2011) is reasonable, the converse is certainly not true. In other words, one may hope that data-driven methods may reach some good sensitivity level for the discovery of non-causal interactions (even if this has not been achieved yet), but expecting good specificity would be illusory. The most important source of non-functionally related relationships may be unmodeled common causes (co-regulators) affecting two variables, inducing a common behavior among these variables that does not correspond to one of them directly or indirectly influencing the other.

Mechanism-driven network inference

What could be the driving principles for the design of mechanistically driven models for interactions among molecular species within a cell? Since unaccounted-for common causes may be seen as the main source of spurious discoveries of relationships, one natural requirement should be to include these causes in the model whenever they are biologically identified. This comes with a price, certainly, creating more complex networks that involve hidden (unobserved) variables. Such networks can then only be identified with drastic constraints on their structures and topology, which is the approach we are recommending, leveraging prior mechanistic knowledge. To be specific, revisit the case of a signaling network, but now include the sequence of intermediate reactions. The signaling proteins are created through biochemical processes captured in the Central Dogma and elucidated over many decades of molecular biology research. The genes encoding the proteins reside in DNA, which are transcribed to RNA, and translated into protein. Transcription is controlled by the transcription factors (TFs) that are downstream effectors of signaling. The TFs transcribe their targets when activated, unless the targets are blocked through methylation of the DNA at promoters of the genes. The genes can also be silenced by compaction of the DNA into chromatin in the region containing the gene. The amount of mRNA produced can also be affected by the copy number, and mRNA may be destroyed if targeted by a micro-RNA (miRNA). The miRNAs are transcribed by TFs as well, with processing through their own cellular machinery to become active. The mRNA for a gene is exported from the nucleus

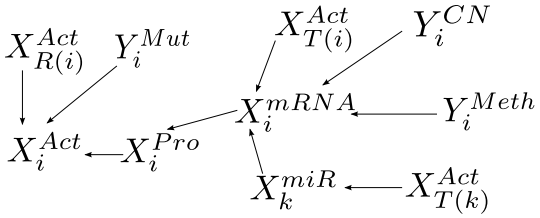


Fig. 4 Expanded gene interaction. *Arrows* indicate the direction of causation between different molecular types, with *subscripts* indexing gene names and *superscripts* indexing molecular type (e.g., mRNA, protein, activated protein, etc.). Some effects are activating and some inhibiting (e.g., methylation). $T(i)$ is the set of transcription factors which regulate gene i, and $R(i)$ is the set of signaling proteins that regulate the activated protein for gene i. The CN (copy number), Meth (methylation) and Mut (mutation) variables are always roots of the network and some mRNA species have an additional hidden variable upstream capturing the expression due to non-modeled components

and translated into protein by ribosomes. As noted above, a signaling protein is inactive until either ligand binding (receptors), post-translation modification such as phosphorylation (signaling proteins), or unless mutated to a constitutively active form.

In a full probabilistic model of signaling, a number of variables must therefore be considered to properly represent all the interactions among genes. Note that some of these variables, like mRNA concentration, would be cell dependent, while others, like copy numbers of genes, are shared among cells, and are essentially constant in a homogeneous tissue. A causal network illustrating this is represented in Fig. 4. In this figure, $X_i^{(*)}$ or $Y_i^{(*)}$ represent variables attached to the gene or protein i, X being used for cell-dependent variables, and Y for tissue-dependent variables that are shared among cells. This graphical model would then need to be nested within the model of Fig. 3. Most recent acquisition tools allow for the observation of an increasing number of these variables, but some still are, and will probably remain in the foreseeable future, unobservable. Moreover, the coexistence of cell-level and tissue-level variables requires that the model be defined at multiple scales, and this is reinforced by the fact that observed data are most of the time aggregated over large numbers of cells within assays (single cell observation being, for the time being, unachievable). The distinction between the statistical model, which is designed at the cell level, with possible tissue-level variables interacting, and the observation, which are tissue-level concentrations, is an important one. The fine analysis of the interactions among molecular species only makes sense at the cell level, and the model of Fig. 4 applied to tissue concentrations would have a very different, and probably inaccurate, interpretation. This, however, comes at a cost, which is that the unobserved variables have a richer structure than the observed one (thousands of cells vs. one aggregate observation), so that the analysis of the model requires using statistical techniques designed for partial observations, combined with strong model assumptions to ensure statistical identifiability. While there has been great progress in single cell measurements, it is likely that data acquisition in many cases will be limited to tissue level measurements comprising collections of cells. Note that the causal structure in Fig. 4 is determined a priori. It is not, and generally cannot be, learned from data.

Mutations and tumorigenesis

Theodor Boveri is credited with formulating the hypothesis that cancer is a genetic disease (Boveri 2008). We now know that cancer is caused by genetic alterations disrupting the function of certain pathways and that the accumulation of these mutational events, known as drivers, is the cause behind the clonal evolution of tumors (Vogelstein et al. 2013). In fact, modern sequencing technologies have permitted the discovery of many of these drivers. Statistical analysis based on probabilistic modeling of somatic mutations' accumulation, cancer initiation and progression are among the most successful examples of the fruitful interaction of probabilistic modeling and statistical analysis with biology. We will briefly mention two examples.

A history of collaboration

The collaboration of Salvador Luria, a microbiologist, with Max Delbrück, a theoretical physicist, resulted in the development of a new statistical analysis (the fluctuation test) to be used on experimental data for testing whether certain genetic mutations in bacteria were the result of selection or rather a random phenomenom occurring in the absence of selection (Luria and Delbrück 1943). The test was based on comparing a Poisson distribution with a novel probabilistic distribution, developed by mathematically modeling the mechanism behind the random acquisition of mutations in bacteria. Their Nobel prize discovery provided the first evidence that bacterial resistance to phages is the result of genetic inheritance caused by random mutations rather than a directed response to selection. Their Ph.D. advisee James Watson, co-discoverer of the double helix structure of DNA with Francis Crick (again a biologist and a biophysicist), describes the summer 3-week long phage course taught by Delbrück as a mathematically oriented approach to biology that constituted "the training ground for many key scientists who laid the foundations of molecular genetics".

About 10 years later, and following some mathematical modeling work by Charles and Luce-Clausen, Fisher and Hollomon (1951), and the statistical analysis of cancer

incidence data on log–log plots by Nordling (1953), the multistage theory of cancer progression was fully established by Peter Armitage, a statistician, and Richard Doll, a physiologist (Armitage and Doll 1954). Armitage and Doll's (1957) main contribution has been to further develop previous work both from a statistical perspective, by considering separately the incidence curves of different types of cancer, as well as from a modeling point of view by dropping the assumption that mutational events are independent, thus considering the exponential growth occurring in subclones possessing fitness advantages. Their work allowed the inference of the required number of rate-limiting steps to cancer. Much research followed their foundational work. Another success of the multistage theory came in 1971 when Alfred Knudson (1971) compared the differences in incidence of retinoblastoma between inherited and non-inherited forms, showing that cancer incidence data provided evidence for two hits required in sporadic retinoblastoma, while the inherited form possessed already one of them. This prediction was later validated experimentally. It is then not surprising that cancer epidemiology tends to be more mathematically grounded than the modeling efforts at the molecular and cellular levels, also due to the contributions of statistical genetics to the field.

The current state: mechanisms and models

The works mentioned above created new research directions in probabilistic modeling of biological systems, especially with regard to the process of tumorigenesis and the development of drug resistance in cancer.

We will start by mentioning the main biological mechanisms that have been included in these models. Peter Nowell (1976) proposed the clonal evolution model of cancer, which was later confirmed by large experimental evidence: cancer typically originates from a single cell, which initiates a clonal expansion where mutational events yield the sequential selection of subclones with increasing fitness advantages thanks to the tumor genetic instability. The occurrence of these mutational events, if not already inherited, may be induced by environmental factors, like carcinogens and viruses, as well as by purely stochastic events, random errors in DNA duplication occurring during a cell division. Similarly, in single or multi-drug resistance, the occurrence of somatic mutations inducing the expansion of clones resistant to a drug appears to be a random phenomenon often not induced by the selective effects of the drug but rather by stochastic events occurred prior to the start of the treatment, as we have already seen in the classical work of Luria and Delbrück. This is particularly true in the case of resistance to the new so-called targeted therapies.

Thus, a large number of stochastic models have been developed in an attempt to characterize the dynamics of tumorigenesis and cancer drug-resistance development, where the mechanisms of random accumulation of mutations and the subsequent cell clonal expansions are included. The literature is too large to mention here in any satisfactory manner, but we will briefly point to a few recent representative examples with the goal of shedding light on the current state of these modeling approaches.

In Durrett and Moseley (2010) the evolution of drug resistance, or alternatively tumor progression, is modeled by an exponentially growing population of wild-type tumor cells, i.e. tumor cells where mutations conferring drug resistance are not present, via a branching process. Subclones of type-i cells, defined as those with $i > 0$ specific mutations, are generated by mutations occurring with rate u_i in the type-$(i-1)$ subpopulation. The needed order of occurrence of the mutations is given and each further subclone is assumed to have a larger fitness (growth) advantage than its immediate predecessor, a possibly limiting element of the model since in the development of drug resistance, mutations may be neutral and even disadvantageous before the start of the drug treatment. Probabilistic techniques via martingales, i.e. stochastic processes whose expected value at the next step is equal to their present value, are then used to derive the distribution for the type-i cell population present at time t and the distribution for the first time at which k mutations have accumulated in some cell.

Some of the limitations in the applicability of this type of mechanism-based probabilistic modeling to experimental data are that the derived closed-form solutions may not be easily tractable statistically and also that the models may not include enough of the biological mechanisms or include them in a simplistic way, for example, by assuming exponential growth of the clonal populations, a requirement probably violated in tumorigenesis given the limited resources present in a tissue and the related concept of a carrying capacity. These types of results have, however, proved to be theoretically useful and, at times, have been used in applied work. For example, in Diaz et al. (2012) a simpler version of the formulas derived from current branching process models is used for the statistical analysis of clinical data to estimate the timing of resistance evolution to targeted EGFR blockade in colorectal cancer, providing evidence in favor of the hypothesis that mutations were already present before the initiation of panitumumab treatment. Beerenwinkel et al. (2007) considers instead the progression of a benign tumor of the colon to a carcinoma, using a Wright–Fisher process with growing population size to estimate the expected waiting time for the tumor to progress from benign to cancer status. The model is also used in conjunction with the statistical analysis of sequencing data of about 13,000 genes, to infer the average selective advantage per driver mutation, finding it to be small (on the order of 1 %). Similarly, Iacobuzio-Donahue and

colleagues (2010) use genome-sequencing data in combination with a Poisson process model to analyze distinct tissue subclones, with the goal of estimating the timescales of the genetic evolution of pancreatic cancer, and inferring that it takes at least 15 years for the tumor initiating mutation to yield a metastatic cancer therefore showing the potential for a useful time-window in detecting cancer at an earlier stage.

The above probabilistic models all consider tumorigenesis at or after the first driver hit, that is, not sooner than the first clonal expansion. Tomasetti et al. (2013) instead investigated the process of accumulation of somatic mutations in a tissue both before and after tumor initiation and progression, estimating the somatic mutation rates in vivo for different human tissues and yielding the unexpected result that even a majority of the mutations found in cancer tissues originates before the process of tumorigenesis initiated. The probabilistic model developed, partially based on Tomasetti and Levy (2010), is an integration of different modeling components for the various phases that a tissue undergoes during its lifespan (development, healthy self-renewal and tumorigenesis). Importantly, while some of the derived formulas are used for statistical inference in combination with exome-sequencing data, the model and its predictions are also used for simply guiding the statistical analysis of the sequencing data, finding age correlations previously not observed.

Thus, the work by Tomasetti et al. emphasizes some of the limits of the current statistical methodologies for addressing problems in cancer genomics like drivers versus passengers identification, number of drivers required by a cancer and so forth. At present, genes are typically called drivers in a simplistic statistical way: if their mutation frequency is larger than expected given some average background rate, which depends on the cancer type (Lawrence et al. 2013). An interesting exception is provided by Vogelstein et al. (2013), where mechanistically based ratiometric scores are used to identify drivers.

Looking forward

Clonal evolution certainly represents a valuable instance of the fruitful interaction of probabilistic modeling and statistical analysis with biology, as indicated, for example, by the success stories we have mentioned. However, we would like to argue for the need of a more extensive use of modeling of biological mechanisms and their temporal dynamics in the analysis of genomics data. This is necessary if we want to both deepen our understanding of the processes analyzed in cancer genomics as well as increase our ability to make risk prediction. Indeed, it is clear from the previous section that the mechanisms included in current models are rather elementary when compared with the complexity of tumorigenesis. While complex models with too many variables will not be statistically useful, there is a need to narrow the gap between models and reality. For example, the current assumption of exponential growth induced by a selective advantage must be modified to allow for the growth rate of those clonal expansions to be a decreasing function of the tumor clone size, when approaching some carrying capacity. Otherwise, the results on the timing of cancer occurrence or on the number of drivers accumulated will be heavily biased.

Moreover, often studies report only one out of many possible evolutionary models, without relying on formal statistical inference methods. Thus, the use of mechanistic-based models together with parsimony assumptions within a more rigorous statistical inference framework is greatly needed in this new era of omics data.

Conclusion

The nearly universal absence of mechanistic underpinnings for the predictors and signatures generated by current statistical learning algorithms represents a crucial barrier toward the successful discovery of novel biology and the implementation of clinically useful biomarkers. “Hard-wiring” potential mechanisms into predictive models is a “win-win”: on the biological side it enhances the translational value of the derived classifiers by hypothesizing causal explanations for disease phenotypes; on the statistical side it forcefully addresses the “curse of dimensionality” by limiting the model space, which increases robustness against overfitting and thereby addresses, in part, the failure of many biomarkers to validate in novel cohorts. Therefore, embedding biological mechanisms into statistical learning has intrinsic added value for knowledge discovery and disease treatment design, and it will ultimately move the field towards a successful transition to personalized health care.

More generally, using prior information to the largest possible extent is a basic principle in statistical modeling which has been somewhat ignored in applications to computational biology even though a large amount of mechanistic biological information is available. This fact can certainly be at least partially explained by the complexity of biological interactions, which makes the construction and learning of adapted statistical models extremely challenging. A second reason may be the optimistic expectation, inspired from striking successes in other areas, like text understanding and pattern recognition, that off-the-shelf data mining methods, independent of prior knowledge, could be applied to high-throughput data and discover new interactions that would be validated a posteriori. In contrast, as we have argued, this approach applied to

computational systems medicine has failed to provide enough reproducible results, compared to the immense effort that has been devoted to it. Finally, another reason, of course, is that working out mechanistically driven statistical models requires a combination of expertise that is rarely achieved in a single individual, and even in a single research group.

Given our goal of identifying mechanistic drivers of tumor growth and metastasis, the use of statistical models that integrate diverse measurements in their biological context is essential. For example, in cell signaling, non-linear effects, such as epistasis, and biological complexity, such as retroactivity, introduce unsuspected mechanisms of response to changes in signaling, whether driven by mutation or targeted therapies. Quantitative models that integrate biological context can address these issues by greatly limiting potential models (e.g., not allowing all gene interactions) while still capturing complex interactions. One example where we have significant mechanistic information on which to leverage is for metabolic networks, which have been mapped out at the genome scale in humans. The potential power of these models has also been demonstrated in studies where identification of a weakness in cancer cell metabolism through biological-informed modeling permitted creation of a targeted therapy. However, the variability of biological systems has tended to limit the value of single therapy approaches, and treatment of this variability (i.e. stochasticity) will be essential to make significant progress. Mechanistic networks such as those in metabolism can also be used to drive 'forward calculations' where predictions for new scenarios can be made from their effects on known mechanisms based on physico-chemical laws, and thus require little to no training, providing another important link between network models and the ability to deal with the enormous complexity and variability of biological systems. Finally, we would like to remark that an under-appreciated use of mechanism-based probabilistic models is to guide the statistical analysis of empirical data, as has been the case in statistical genetics.

In summary, statistical methods based on probabilistic modeling have yielded fundamental contributions to biology. It can be claimed that those contributions are the consequence of formulating probabilistic models of specific biological mechanisms, that is mechanism-based models, which are then used for the statistical analysis of experimental, clinical and epidemiological data. At the same time much of the work in mathematical modeling and statistical analysis has suffered for the lack of statistical tractability in the former case and model naivety in the latter case, failing to provide clinically relevant inference and risk prediction in cancer biology. The aim is to strike the right balance between models, their statistical analysis and the experiments, a fact that highlights the need for true collaborations and researchers well versed across those fields. Developing deeper interactions between cutting-edge statistics and biology is one of the challenges of research in computational biology in the years to come.

Acknowledgments The work of D. Geman and L. Younes was partially supported by the National Science Foundation under NSF DMS1228248. N. Price's work was supported by a Camille Dreyfus Teacher-Scholar Award and NIH 2P50GM076547.

Author contributions D.G. supervised the project. All authors wrote the manuscript.

References

Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J (2010) Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinform 11:277. doi:10.1186/1471-2105-11-277

Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. PLoS Comput Biol 8(5):e1002518. doi:10.1371/journal.pcbi.1002518

Altman R (2012) Translational bioinformatics: linking the molecular world to the clinical world. Clin Pharmacol Ther 91(6):994–1000

Altman RB, Kroemer Ho K, McCarty CA et al (2011) Pharmacogenomics: will the promise be fulfilled. Nat Rev 12:69–73

Anderson AR, Tomlin CJ, Couch J, Gallahan D (2013) Mathematics of the integrative cancer biology program. Interface Focus 3(4):20130023

Armitage P, Doll R (1954) The age distribution of cancer and a multistage theory of carcinogenesis. Br J Cancer 8(1):1–12. URL http://www.ncbi.nlm.nih.gov/pubmed/13172380

Armitage P, Doll R (1957) A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. Br J Cancer 11(2):161–169. URL http://www.ncbi.nlm.nih.gov/pubmed/13460138

Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med 1(1):2

Barrett CL, Price ND, Palsson BO (2006) Network-level analysis of metabolic regulation in the human red blood cell using random sampling and singular value decomposition. BMC Bioinform 7:132. doi:10.1186/1471-2105-7-132

Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, Vogelstein B, Nowak MA (2007) Genetic progression and the waiting time to cancer. PLoS Comput Biol 3(11):e225. doi:10.1371/journal.pcbi.0030225. URL http://www.ncbi.nlm.nih.gov/pubmed/17997597

Bender R, Knauer M, Rutgers E, Glas A, de Snoo FA et al (2009) The 70-gene profile and chemotherapy benefit in 1,600 breast cancer patients. J Clin Oncol 27(18 Suppl):512

Binder H, Schumacher M (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinform 10:18. doi:10.1186/1471-2105-10-18

Bordbar A, Lewis NE, Schellenberger J, Palsson BØ, Jamshidi N (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. Mol Syst Biol 6:422. doi:10.1038/msb.2010.68

Boulesteix AL, Sauerbrei W (2011) Added predictive value of high-throughput molecular data to clinical data and its validation. Brief Bioinform 12(3):215–229

Boulesteix AL, Tutz G, Strimmer K (2003) A cart-based approach to discover emerging patterns in microarray data. Bioinformatics 19(18):2465–2472

Boveri T (2008) Concerning the origin of malignant tumours by theodor boveri. translated and annotated by henry harris. J Cell Sci 121(Suppl 1):1–84. doi:10.1242/jcs.025742. URL http://www.ncbi.nlm.nih.gov/pubmed/18089652

Brenner S (2010) Sequences and consequences. Philos Trans R Soc Lond B Biol Sci 365(1537):207–212

Butte AJ (2008) Translational bioinformatics: coming of age. J Am Med Inform Assoc 15(6):709–714

Butte AJ, Kohane IS (2003) Relevance networks: a first step toward finding genetic regulatory networks within microarray data. In: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (eds) The analysis of gene expression data, pp 428–446

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci 97(22):12182–12186

Chandrasekaran S, Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 107(41):17845–17850. doi:10.1073/pnas.1005139107

Chandrasekaran S, Price ND (2013) Metabolic constraint-based refinement of transcriptional regulatory networks. PLoS Comput Biol 9(12):e1003370. doi:10.1371/journal.pcbi.1003370

Chavali AK, Whittemore JD, Eddy JA, Williams KT, Papin JA (2008) Systems analysis of metabolism in the pathogenic trypanosomatid leishmania major. Mol Syst Biol 4:177. doi:10.1038/msb.2008.15. URL http://www.ncbi.nlm.nih.gov/pubmed/18364711

Chen X, Wang L, Ishwaran H (2010) An integrative pathway-based clinical-genomic model for cancer survival prediction. Stat Probab Lett 80(17–18):1313–1319. doi:10.1016/j.spl.2010.04.011

Cohen JE (2004) Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. PLoS Biol 2(12):e439

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987):92–96. doi:10.1038/nature02456

Croce CM (2009) Causes and consequences of microrna dysregulation in cancer. Nat Rev Genet 10(10):704–714. doi:10.1038/nrg2634

Cronin M, Sangli C, Liu ML, Pho M, Dutta D, Nguyen A, Jeong J, Wu J, Langone KC, Watson D (2007) Analytical validation of the oncotype dx genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. Clin Chem 53(6):1084–1091

Dettling M, Buhlmann P (2003) Boosting for tumor classification with gene expression data. Bioinformatics 19(9):1061–1069. URL http://www.ncbi.nlm.nih.gov/pubmed/12801866

Diaz LA, Williams RT, Wu J, Kinde I, Hecht JR, Berlin J, Allen B, Bozic I, Reiter JG, Nowak MA, Kinzler KW, Oliner KS, Vogelstein B (2012) The molecular evolution of acquired resistance to targeted egfr blockade in colorectal cancers. Nature 486(7404):537–540. doi:10.1038/nature11219. URL http://www.ncbi.nlm.nih.gov/pubmed/22722843

Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97(457):77–87

Durrett R, Moseley S (2010) Evolution of resistance and progression to disease during clonal expansion of cancer. Theor Popul Biol 77(1):42–48. doi:10.1016/j.tpb.2009.10.008. URL http://www.ncbi.nlm.nih.gov/pubmed/19896491

Eddy JA, Hood L, Price ND, Geman D (2010) Identifying tightly regulated and variably expressed networks by differential rank conservation (dirac). PLoS Comput Biol 6(5):e1000792. doi:10.1371/journal.pcbi.1000792

Edelman LB, Toia G, Geman D, Zhang W, Price ND (2009) Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. BMC Genomics. doi:10.1186/1471-2164-10-583

Eisen MB, Spellman PT, Brown PO (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95(25):14863–14868

Evans JP, Meslin EM, Marteau TM, Caulfield T (2011) Deflating the genomic bubble. Science 331:861–862

Fisher JC, Hollomon JH (1951) A hypothesis for the origin of cancer foci. Cancer 4(5):916–918. URL http://www.ncbi.nlm.nih.gov/pubmed/14879355

Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7:501. doi:10.1038/msb.2011.35

Frezza C, Zheng L, Folger O, Rajagopalan KN, MacKenzie ED, Jerby L, Micaroni M, Chaneton B, Adam J, Hedley A, Kalna G, Tomlinson IPM, Pollard PJ, Watson DG, Deberardinis RJ, Shlomi T, Ruppin E, Gottlieb E (2011) Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. Nature 477(7363):225–228. doi:10.1038/nature10363

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303(5659):799–805

Geman D, d'Avignon C, Naiman D et al (2004) Gene expression comparisons for class prediction in cancer studies. In: Proceedings 36th symposium on the interface: computing science and statistics

Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias variance dilemma. Neural Comput 4(1):1–58

Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100(1):57–70

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674. doi:10.1016/j.cell.2011.02.013

Hartemink AJ et al (2005) Reverse engineering gene regulatory networks. Nat Biotechnol 23(5):554–555

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Hobert O (2008) Gene regulation by transcription factors and microrna. Science 319(5871):1785–1786. doi:10.1126/science.1151651

Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–643

Hood L, Price ND (2014) Demystifying disease, democratizing health care. Sci Transl Med 6(225):225ed5

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2(1):343–372

Jamshidi N, Palsson BO (2006) Systems biology of SNPs. Mol Syst Biol 2:38

Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of mycobacterium tuberculosis h37rv using the in silico strain inj661 and proposing alternative drug targets. BMC Syst Biol 1:26. doi:10.1186/1752-0509-1-26. URL http://www.ncbi.nlm.nih.gov/pubmed/17555602

Jerby L, Ruppin E (2012) Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. Clin Cancer Res 18(20):5572–5584. doi:10.1158/1078-0432.CCR-12-1856

Johannes M, Brase JC, Fröhlich H, Gade S, Gehrmann M, Fälth M, Sültmann H, Beissbarth T (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics 26(17):2136–2144. doi:10.1093/bioinformatics/btq345

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at genomenet. Nucleic Acids Res 30(1):42–46

Kern SE (2012) Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. Cancer Res 72(23):6097–6101. doi:10.1158/0008-5472.CAN-12-3232

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C et al (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7(6):673–679

Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8(2):1002. doi:10.1371/Journal.Pcbi375

Kim YA, Wuchty S, Przytycka TM (2011) Identifying causal genes and dysregulated pathways in complex diseases. PLOS Comput Biol 7(3):e1001095

Knudson AG (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA 68(4):820–823 (1971). URL http://www.ncbi.nlm.nih.gov/pubmed/5279523

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge

Kreeger PK, Lauffenburger DA (2010) Cancer systems biology: a network modeling perspective. Carcinogenesis 31(1):2–8

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457):214–218 (2013). doi:10.1038/nature12213. URL http://www.ncbi.nlm.nih.gov/pubmed/23770567

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I et al (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799–804

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11(10):733–739. doi:10.1038/nrg2825

Levy R, Borenstein E (2013) Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci USA 110(31):12,804–12,809

Li C, Li H (2008) Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24(9):1175–1182. doi:10.1093/bioinformatics/btn081

Li Q, Seo JH, Stranger B, McKenna A, Pe'er I, Laframboise T, Brown M, Tyekucheva S, Freedman ML (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152(3):633–641. doi:10.1016/j.cell.2012.12.034

Li XJ, Hayward C, Fong PY, Dominguez M, Hunsucker SW, Lee LW, McLean M, Law S, Butler H, Schirm M, Gingras O, Lamontagne J, Allard R, Chelsky D, Price ND, Lam S, Massion PP, Pass H, Rom WN, Vachani A, Fang KC, Hood L, Kearney P (2013) A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. Sci Transl Med 5(207):207ra142. doi:10.1126/scitranslmed.3007013. URL http://www.ncbi.nlm.nih.gov/pubmed/24132637

Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM (2012) Identifying dysregulated pathways in cancers from pathway interaction networks. BMC Bioinform 13:126

Liu Y, Koyuturk M, Barnholtz-Sloan JS, Chance MR (2012) Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. BMC Syst Biol 6:65

Lottaz C, Spang R (2005) Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. Bioinformatics 21(9):1971–1978. doi:10.1093/bioinformatics/bti292

Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28(6):491–511. URL http://www.ncbi.nlm.nih.gov/pubmed/17247100

Maathuis MH, Colombo D, Kalisch M, Bühlmann P (2010) Predicting causal effects in large-scale systems from observational data. Nat Methods 7(4):247–248

Maathuis MH, Kalisch M, Bühlmann P et al (2009) Estimating high-dimensional intervention effects from observational data. Ann Stat 37(6A):3133–3164

Marchionni L, Wilson RF, Wolff AC, Marinopoulos S, Parmigiani G, Bass EB, Goodman SN (2008) Systematic review: gene expression profiling assays in early stage breast cancer. Ann Intern Med 148(5):358–369

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform 7(Suppl 1):S7

Mendell JT (2005) Micrornas: critical regulators of development, cellular physiology and malignancy. Cell Cycle 4(9):1179–1184

Milne CB, Kim PJ, Eddy JA, Price ND (2009) Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. Biotechnol J 4(12):1653–1670. doi:10.1002/biot.200900234

Neapolitan RE et al (2004) Learning bayesian networks, vol 1. Prentice Hall, Upper Saddle River

Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM (2012) Paradigm-shift predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 28(18):i640–i646. doi:10.1093/bioinformatics/bts40210.1093/bioinformatics/bts402. URL http://www.ncbi.nlm.nih.gov/pubmed/22962493.

Nordling CO (1953) A new theory on cancer-inducing mechanism. Br J Cancer 7(1):68–72 (1953). URL http://www.ncbi.nlm.nih.gov/pubmed/13051507

Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194(4260):23–28. URL http://www.ncbi.nlm.nih.gov/pubmed/959840

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320. doi:10.1038/msb.2009.77. URL http://www.ncbi.nlm.nih.gov/pubmed/19888215

Ochs MF, Farrar JE, Considine M, Wei Y, Meshinchi S, Arceci RJ (2014) Outlier analysis and top scoring pair for integrated data analysis and biomarker discovery. IEEE/ACM Trans Comput Biol Bioinform. doi:DBACF900-6B21-49D2-9D30-F333A1E9CED0

Ochs MF, Rink L, Tarn C, Mburu S, Taguchi T, Eisenberg B, Godwin AK (2009) Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res 69(23):9125–9132

Omenn G, DeAngelis C, DeMets D, Fleming T, Geller G, Gray J, Hayes D, Henderson C, Kessler L, Lapidus S, Leonard D, Moses H, Pao W, Pentz R, Price ND, Quackenbush J, Railey E,

Ransohoff D, Reese E, Witten D (2012) Evolution of translational omics: lessons learned and the path forward. Institute of Medicine Report

Paik S (2011) Is gene array testing to be considered routine now? Breast 20(Suppl 3):S87–S91. doi:10.1016/S0960-9776(11)70301-0

Pan W, Xie B, Shen X (2010) Incorporating predictor network in penalized regression with application to microarray data. Biometrics 66(2):474–484. doi:10.1111/j.1541-0420.2009.01296.x

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA Jr, Hartigan J, Smith DR, Strausberg RL, Marie SKN, Shinjo SMO, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW (2008) An integrated genomic analysis of human glioblastoma multiforme. Science 321(5897):1807–1812. doi:10.1126/science.1164382

Patnaik SK, Kannisto E, Knudsen S, Yendamuri S (2010) Evaluation of microrna expression profiles that may predict recurrence of localized stage i non-small cell lung cancer after surgical resection. Cancer Res 70(1):36–45

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo

Pearl J (2000) Causality: models, reasoning and inference, vol 29. Cambridge University Press, Cambridge

Pe'er D, Hacohen N (2011) Principles and strategies for developing network models in cancer. Cell 144(6):864–873

Porzelius C, Johannes M, Binder H, Beissbarth T (2011) Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients. Biom J 53(2):190–201. doi:10.1002/bimj.201000155

Price ND, Reed JL, Palsson BØ (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11):886–897. doi:10.1038/nrmicro1023

Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, Hood L, Shmulevich I, Zhang W (2007) Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proc Natl Acad Sci USA 104(9):3414–3419. doi:10.1073/Pnas.0611373104

Raponi M, Lancet JE, Fan H, Dossey L, Lee G, Gojo I, Feldman EJ, Gotlib J, Morris LE, Greenberg PL, Wright JJ, Harousseau JL, Lowenberg B, Stone RM, De Porre P, Wang Y, Karp JE (2008) A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia. Blood 111(5):2589–2596. doi:10.1182/blood-2007-09-112730. URL http://www.ncbi.nlm.nih.gov/pubmed/18160667

Rejniak KA, Anderson AR (2012) State of the art in computational modeling of cancer. Math Med Biol 29(1):1–2

Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721):523–529

Schadt EE, Björkegren JLM (2012) New: network-enabled wisdom in biology, medicine, and health care. Sci Transl Med 4(115):115rv1. doi:10.1126/scitranslmed.3002132

Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E (2011) Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. PLoS Comput Biol 7(3):e1002018. doi:10.1371/journal.pcbi.1002018

Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26(9):1003–1010. doi:10.1038/nbt.1487

Simcha DM, Younes L, Aryee MJ, Geman D (2013) Identification of direction in gene networks from expression and methylation. BMC Syst Biol 7(1):118

Simon R (2006) Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. J Natl Cancer Inst 98(17):1169–1171. doi:10.1093/jnci/djj364

Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95(1):14–18

Staiger C, Cadot S, Kooter R, Dittrich M, Müller T, Klau GW, Wessels LFA (2012) A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PLoS One 7(4):e34796. doi:10.1371/journal.pone.0034796

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550. doi:10.1073/pnas.0506580102

Sung J, Kim PJ, Ma S, Funk CC, Magis AT, Wang Y, Hood L, Geman D, Price ND (2013) Multi-study integration of brain cancer transcriptomes reveals organ-level diagnostic signatures. PLoS Comput Biol 9(7):e1003148

Sung J, Wang Y, Chandrasekaran S, Witten DM, Price ND (2012) Molecular signatures from omics data: from chaos to consensus. Biotechnol J 7(8):946–57. doi:10.1002/biot.201100305. URL http://www.ncbi.nlm.nih.gov/pubmed/22528809

Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D (2005) Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21(20):3896–3904 (2005). doi:10.1093/bioinformatics/bti631. URL http://www.ncbi.nlm.nih.gov/pubmed/16105897

Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bölling C, Bordel S, Chavali AK, Dobson P, Dunn WB, Endler L, Hala D, Hucka M, Hull D, Jameson D, Jamshidi N, Jonsson JJ, Juty N, Keating S, Nookaew I, Le Novère N, Malys N, Mazein A, Papin JA, Price ND, Selkov E Sr, Sigurdsson MI, Simeonidis E, Sonnenschein N, Smallbone K, Sorokin A, van Beek JHGM, Weichart D, Goryanin I, Nielsen J, Westerhoff HV, Kell DB, Mendes P, Palsson BØ (2013) A community-driven global reconstruction of human metabolism. Nat Biotechnol 31(5):419–425. doi:10.1038/nbt.2488

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99(10):6567–6572

Tomasetti C, Levy D (2010) An elementary approach to modeling drug resistance in cancer. Math Biosci Eng 7(4):905–918. URL http://www.ncbi.nlm.nih.gov/pubmed/21077714

Tomasetti C, Vogelstein B, Parmigiani G (2013) Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci USA 110(6):1999–2004. doi:10.1073/pnas.1221068110. URL http://www.ncbi.nlm.nih.gov/pubmed/23345422

Tuncbag N, Braunstein A, Pagnani A, Huang SS, Chayes J, Borgs C, Zecchina R, Fraenkel E (2013) Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. J Comput Biol 20(2):124–36. doi:10.1089/cmb.2012.0092. URL http://www.ncbi.nlm.nih.gov/pubmed/23383998.

Ulitsky I, Krishnamurthy A, Karp RM, Shamir R (2010) Degas: de novo discovery of dysregulated pathways in human diseases. PLoS One 5(10):e13367

Vandin F, Clay P, Upfal E, Raphael B (2012) Discovery of mutated subnetworks associated with clinical data in cancer. In: Proceedings Pacific symposium biocomputing, pp 55–66

Varadan V, Mittal P, Vaske CJ, Benz SC (2012) The integration of biological pathway knowledge in cancer genomics: a review of existing computational approaches. IEEE Signal Process Mag 29(1):35–50. doi:10.1109/Msp.2011.943037

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu JC, Haussler D, Stuart JM (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26(12):i237–i245. doi:10.1093/bioinformatics/btq182

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW (2013) Cancer genome landscapes. Science 339(6127):1546–1558. doi:10.1126/science.1235122. URL http://www.ncbi.nlm.nih.gov/pubmed/23539594

Wang Y, Eddy JA, Price ND (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol 6(1):153. doi:10.1186/1752-0509-6-153

Wei Z, Li H (2007) Non-parametric pathway-based regression models for analysis of genomic data. Biostatistics 8(2):265–284. doi:10.1093/biostatistics/kxl007

Weichselbaum RR, Ishwaran H, Yoon T, Nuyten DSA, Baker SW, Khodarev N, Su AW, Shaikh AY, Roach P, Kreike B, Roizman B, Bergh J, Pawitan Y, de Vijver MJV, Minn AJ (2008) An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. Proc Natl Acad Sci USA 105(47):18490–18495. doi:10.1073/Pnas.0809242105

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113–1120. doi:10.1038/ng.2764

Wilson JL, Hemann MT, Fraenkel E, Lauffenburger DA (2013) Integrated network analyses for functional genomic studies in cancer. Semin Cancer Biol 23(4):213–218. doi:10.1016/j.semcancer.2013.06.004. URL http://www.ncbi.nlm.nih.gov/pubmed/23811269.

Winslow R, Trayanova N, Geman D, Miller M (2012) The emerging discipline of computational medicine. Science Transl Med 4(158):158rv11

Winslow RL, Trayanova N, Geman D, Miller MI (2012) Computational medicine: translating models to clinical care. Sci Transl Med 4(158):158rv11. doi:10.1126/scitranslmed.3003528

Wynn ML, Ventura AC, Sepulchre JA, García HJ, Merajver SD (2011) Kinase inhibitors can produce off-target effects and activate linked pathways by retroactivity. BMC Syst Biol 5:156. doi:10.1186/1752-0509-5-156

Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics 21(20):3905–3911. doi:10.1093/bioinformatics/bti647

Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, Kamiyama M, Hruban RH, Eshleman JR, Nowak MA, Velculescu VE, Kinzler KW, Vogelstein B, Iacobuzio-Donahue CA (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467(7319):1114–1117. doi:10.1038/nature09515. URL http://www.ncbi.nlm.nih.gov/pubmed/20981102

Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, Reich M, Lander E, Mesirov J, Golub T (2001) Molecular classification of multiple tumor types. Bioinformatics 17(Suppl 1):S316–S322. URL http://www.ncbi.nlm.nih.gov/pubmed/11473023

Yoruk E, Ochs MF, Geman D, Younes L (2011) A comprehensive statistical model for cell signaling. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 8(3):592–606

Zhang D, Tai LK, Wong LL, Chiu LL, Sethi SK, Koay ES (2005) Proteomic study reveals that proteins involved in metabolic and detoxification pathways are highly expressed in her-2/neu-positive breast cancer*. Mol Cell Proteomics 4(11):1686–1696

Zhao H, Logothetis CJ, Gorlov IP (2010) Usefulness of the top-scoring pairs of genes for prediction of prostate cancer progression. Prostate Cancer Prostateic Dis 13(3):252–259 (2010). doi:10.1038/pcan.2010.9. URL http://www.ncbi.nlm.nih.gov/pubmed/20386565

Zhu Y, Shen X, Pan W (2009) Network-based support vector machine for classification of microarray samples. BMC Bioinform 10(Suppl 1):S21. doi:10.1186/1471-2105-10-S1-S21