Published in final edited form as:

Prog Biophys Mol Biol. 2011 August ; 106(2): 410–413. doi:10.1016/j.pbiomolbio.2011.03.002.

Systems Biology and Cancer: Promises and Perils

Stuart G. Baker and

Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, U.S.A.

Barnett S. Kramer

Office of Communications and Education, National Cancer Institute (contractor), Bethesda, MD, U.S.A.

Abstract

Systems biology uses systems of mathematical rules and formulas to study complex biological phenomena. In cancer research there are three distinct threads in systems biology research: modeling biology or biophysics with the goal of establishing plausibility or obtaining insights, modeling based on statistics, bioinformatics, and reverse engineering with the goal of better characterizing the system, and modeling with the goal of clinical predictions. Using illustrative examples we discuss these threads in the context of cancer research.

Keywords

bioinformatics; microarrays; reverse engineering; receiver operating characteristic curves

The term “systems biology” has different meanings for different groups of researchers. One view is that “systems biology [is]... a powerful new paradigm... based on the premise that the properties of complex systems consisting of many components that interact with each other in non-linear, non-additive ways cannot be understood solely by focusing on the components’ (Laubenbacher et al, 2008). A reviewer of this manuscript defined systems biology as “an interdisciplinary study of biological systems that spans multiple scales in space and time.” In our view, systems biology is the study of biological phenomena using systems of mathematical rules or formulas.

There is a large literature involving systems biology and actual or potential applications to cancer research (e.g., the review article by Laubenbacher et. al., 2009, and other references in that paper and here). What is sometimes missing is a perspective that sees various threads of systems biology as different parts of a single fabric. In particular, we see a divide between systems biology based on the principles of biology or biophysics, systems biology related to statistics, bioinformatics, and reverse engineering, and systems biology involving clinical predictions, sometimes without full appreciation of other viewpoints. To provide perspective, we briefly discuss these three major threads in system biology and their relation to cancer research. This is not a comprehensive review, but rather draws upon selected articles as illustrative examples.

Corresponding author: Stuart G. Baker, Sc.D. National Cancer Institute, EPN 3131, 6130 Executive Blvd MSC 7354 Bethesda, MD 20892-7354 Phone: (301) 496-7708 sb16i@nih.gov Fax: (301) 402-0816 <http://www.cancer.gov/prevention/bb/baker.html>.

The opinions in this manuscript are those of the authors and do not necessarily represent official opinions or positions of the National Institutes of Health, the National Cancer Institute, or the Department of Health and Human Services.

1. Establishing plausibility and gaining insight

In one strand of systems biology, the goal is to investigate the plausibility of the systems model and sometimes provide new insights into the underlying biologic process. Often there is no formal attempt to estimate parameters from the data. Instead investigators postulate a mathematical model and parameter values, and then run the model to see if it qualitatively mimics observed phenomenon. Alternatively, or in addition, investigators derive properties of the mathematical model that might lead to new insights. By its nature, this form of systems biology is exploratory. However, it can illuminate problems or paradoxes in the current dominant explanations of biological observations and sometimes provide plausible alternative hypotheses. Although this approach can provide challenges to currently accepted paradigms, it can also be criticized as speculation without experimental proof. However it is an important early step in the scientific method—the attempt to advance our understanding whenever empirically observed paradoxes surface.

One example of this use of systems biology is a comprehensive mathematical construct for the development of cancer in which Ao et al (2008) hypothesized that “molecular and cellular agents, such as oncogenes and suppressor genes, and related growth factors hormones, cytokines, etc, form a nonlinear, stochastic, and collective dynamic network”. Although Ao et al (2008) did not provide a detailed diagram “because the precise information ... remains to be obtained”, they listed possible components of a model for the development of prostate cancer. Based on a general mathematical framework, Ao et al (2008) postulated that healthy tissue corresponds to a globally stable mathematical state and neoplastic dormancy corresponds to a locally stable mathematical state.

A different example of this strand of systems biology is an investigation into how cells assemble into tissue. Lemon et al (2006) formulated a model for tissue growth based on mechanistic equations for fluid flow. The mathematical results agreed qualitatively with experimental results involving the seeding of cells onto various surfaces. A main result was that cells either formed aggregates or were uniformly diffused through a biological scaffold.

Yet another example of this use of systems biology is a study of morphostats, which are substances thought to diffuse within tissue to maintain tissue organization. Some investigators believe that the disruption of morphostats could lead to carcinogenesis (Soto and Sonnenschein, 2004; Potter, 2007, Sonnenschein and Soto, 2008). To mathematically investigate how the blockage of morphostats could relate to early-stage cancer, Baker et al (2009) created a rule-based model of the natural cycle of cell regeneration in epithelial tissue superimposed on a morphostat gradient that induces cell differentiation. Baker et al (2009) demonstrated that their model of morphostat diffusion could reproduce the experimental result that more tumors arise when support tissue is blocked by a filter with small holes than when blocked by a filter with large holes. Importantly the model did not rely upon gene mutations to explain observed phenomena in carcinogenesis.

The aforementioned model of morphostats is an example from the more general class of agent-based models for systems biology. Agent-based models are rule-based models incorporating components (agents) that evolve in discrete time units, and have been developed for a wide variety of biological systems, for example the effect of inflammation on hepatocellular carcinoma (An, et al, 2009).

A general caution when drawing conclusions from these models is that they do not rule out the possibility that other models could mimic the same phenomenon. However, if investigators can formulate a few plausible models that yield similar results, they may be able to design experiments to collect data that can distinguish among these models. These

models can serve as good basis for further research and, if correct, provide insight that could be empirically supported or refuted.

2. Inferring systems components from gene expression data

A more ambitious goal of systems biology is to try to identify the system components and their connections, a procedure sometimes called reverse engineering (Wang et al, 2007). Many of these approaches are based on gene expression data. Various reverse engineering algorithms have been proposed, including mathematical models based on either molecular kinetics (Aijo and Landesmaki, 2009) or statistical associations (Werhli et al, 2006). The three primary types of statistical associations in these models take the form of either relevance networks, in which network connections are identified from pair-wise associations, graphical Gaussian models, in which network connections are identified from an estimated variance-covariance matrix, or Bayesian networks, in which network connections are identified from conditional distributions (Werhli et al 2006).

A reasonable measure of how correctly the network is specified is a receiver-operating characteristic (ROC) curve. The ROC curve plots true versus false positive rates. In the context of network models, the false positive rate is the fraction of non-existent connections incorrectly specified by the model as existent, and the true positive rate is the fraction of existent connections correctly specified by the model as existent. An indication of good network specification by reverse engineering is high true positive rate and a low false positive rate in simulated data from a known network. Based on simulated data, reverse engineering has achieved good network specification for small networks, such as those involving 5 components (Aijo and Landesmaki, 2009) or 11 components (Werhli et al 2006).

An open question is whether or not reverse engineering can successfully identify networks using data on gene expression microarrays for cancer, which involve thousands of genes. We identify the following five challenges in this regard.

(1) Large sparse networks

Genetic networks from larger studies are likely to be sparse, meaning that many nodes are not directly connected. Margolin et al (2007) noted that with sparse networks, the false positive rate is high relative to the true positive rate. Margolin et al (2007) concluded that the false positive rate is an inappropriate statistic for the evaluation of sparse networks. However we disagree with this assessment of the inappropriateness of the false positive rate in evaluation of sparse networks. The false and true positive rates are standard measures that are used together to summarize classification performance. (Use of one without the other is not meaningful). In evaluation one needs to weigh the consequence of a false positive against the benefit of a true positive. This weighing of consequences does not depend on the type of data, and in particular does not depend on whether or not a network is sparse. False leads generate expenditure of wasted time and resources irrespective of the field of scientific investigation.

(2) Measurement error

In the context of gene expression microarrays, measurement error refers to the variability in gene expression levels that are measured in replicates from the same specimen. This measurement error of gene expression levels can arise from variability in hybridization, background fluorescence, or signal quantification (Fujita et al, 2009). Using a simulation study, Fujita et al (2009) concluded that “measurement error dangerously affects the identification of regulatory network models, thus they must be reduced or taken into account in order to avoid erroneous conclusions. This could be one of the reasons for the high

biological false positive rates identified in actual regulatory network models.” There may also be a more fundamental measurement problem with gene expression microarrays, namely that they do not accurately measure the abundance of mRNA because they do not account for mRNA decay rate (Rosenfeld, 2009).

(3) Chaotic regimes

On the basis of systems of equations for biological processes, Rosenfeld (2009) concluded that a stable state would take the form of limit cycle (oscillations) or a chaotic attractor. Under chaos theory even near-perfect measurement and model specification may not yield accurate information about a complex genetic network, although it may be possible to identify some highly predictive genes (Baker and Kramer, 2006). By analogy, while it is possible to obtain good predictions of seasonal climate, daily weather is inherently difficult to predict due to its chaotic nature.

(4) Lack of statistical reproducibility

A well established method of evaluating classification models for microarrays is to randomly split the data into a training and test sample, fit the model into the training sample and evaluate classification performance in the test sample. Empirically, however, when data sets on gene expression microarrays and cancer have been randomly split many different ways into training and test samples, there is great variability in the genes selected for the classification rule in the training sample (Michiels et al, 2005, Baker and Kramer, 2006, Pittelkow and Wilson, 2009). Because of this variability Baker and Kramer (2006) focused on identifying the most frequently occurring genes. To investigate the statistical reproducibility of genetic networks, systems biologists should investigate the variability of networks using this procedure of multiple random splits into training and test samples. Because networks are more complex than individual genes, there will likely be even more variability in network specification than gene specification over these random splits. Also the final test of whether the network is correctly specified would have to come from its study in separate populations—which is a greater challenge. Whether or not it is possible to distinguish “signal” from “noise” in specifying these networks is a matter for future research.

(5) Biological assumptions

Wang et al (2007) motivated the application of systems biology of genetic networks to cancer with the assumption that “the accumulation of genetic mutations in part triggers tumor development and progression”, adding that “because gene activity and regulation define a cancer phenotype it is essential to have a comprehensive understanding of the precise genetic mutations and consequences of these mutations and genetic alterations.” The paradigm described by Wang et al (2007) is a corollary of the dominant somatic mutation theory. However, based on experimental evidence involving foreign-body carcinogenesis and transplantation experiments as well as various observational studies, some investigators have questioned the somatic mutation theory (Soto and Sonnenschein, 2004; Baker and Kramer, 2007a, Potter, 2007, Sonnenschein and Soto, 2008, Baker et al 2009). The core theory must be correct if investigations based on the theory are to yield reproducible results.

3. Predicting cancer outcome based on systems models derived from microarrays

Ultimately, one would like to use systems biology to better predict cancer outcome or to guide and individualize therapy. Better prediction of cancer outcome could lead to more

focused randomized trials of patients at higher risk or most likely to benefit from a new experimental therapy, as determined by a systems biology model.

We identify three general strategies for formulating models to predict cancer outcomes using systems biology. One strategy involves first estimating components and connections using the reverse engineering approach, but this involves the challenges previously discussed. A second strategy is to group expression levels by genetic pathway, but this is not feasible because the majority of human genes have not been assigned to a definitive or specific pathway (Taylor et al, 2009). A third potentially fruitful, strategy is to identify genetic networks from protein networks, assign scores to these networks, and substitute these scores into a prediction model, such as logistic regression (Chuang et al, 2007, Taylor et al, 2009). The prediction models are formulated in a training sample and evaluated in a test sample.

A standard method for evaluating prediction models is an ROC curve that plots false and true positive rates for various classification rules determined by different cutpoints (e.g. Baker, 2003). For these prediction models, the false positive rate is the fraction of negative outcomes incorrectly predicted as positive, and the true positive rate is the fraction of positive outcomes correctly predicted as positive. A well-known result in medical decision-making is that the slope of the ROC curve equals the probability of a negative outcome divided by the probability of a positive outcome multiplied by the cost of a false positive divided by the benefit of a true positive (e.g. Baker and Kramer, 2007b). Going from left to right on the ROC curve, the slope decreases (Figure 1), which corresponds to smaller ratios of the cost of a false positive to the benefit of a true positive, for a fixed probability of disease. When using an ROC curve for evaluation, one should focus on the part of the curve of interest based on the target value for the slope as determined by the aforementioned costs and benefits and the probability of disease.

To put this evaluation of prediction models in perspective, it is important to compare predictions from system biology models with predictions from models with clinical covariates. This is where collaboration between statisticians and clinicians is particularly important. Also when the ultimate goal is clinical utility, one can gain perspective relative to perfect prediction by using relative utility curves instead of ROC curves (Baker, 2009).

As one example, Chuang et al (2007) used this strategy of deriving gene networks from protein networks to predict metastatic versus non-metastatic breast cancer in two studies. For the two studies, Chuang et al (2007) obtained true positive rates of 49% and 56%, corresponding to false positive rates of 10%. In comparison, for single gene markers with the two studies, Chuang et al (2007) reported true positive rates of 45% and 41% corresponding to false positive rates of 10%. Similarly, Taylor et al (2009) used this strategy to predict good or poor outcomes in breast cancer patients. Taylor et al (2009) computed approximately the same true positive rate of 30% for predictions based on networks, clinical covariates, and the combination of the two, corresponding to a false positive rate of 10% (as read from Figure 3a in [17]). This thread of systems biology is theoretically appealing. However given that the “oncology literature is replete with publications on prognostic factors but very few of these are used in clinical practice” (Simon, 2008), considerable more work will likely be needed before these models are clinically useful. Nevertheless it is a very fruitful area for further research.

5. Conclusion

Applications of systems biology to cancer involve both promises and perils, which is the underlying thesis of this commentary.

The great promise of systems biology comes from idea that studying a system can provide information not available by separately studying the workings of each part. As DJ Smithers (1962) wrote “Cancer is no more a disease of cells than a traffic jam is a disease of cars. A lifetime of study of the internal-combustion engine would not help anyone to understand our traffic problems”. The peril comes when the rules leading to a complex system vary over many components and the sample sizes are limited for identifying the rules and making predictions. Suppose that on a highway nearly saturated with cars, a traffic jam arises when a cautious driver slows down when tailgated by an aggressive driver. When viewing a random sample of cars on the highway, it may be difficult to identify the cautious and aggressive drivers unless they happen to be near each other (in which case the traffic jam has already started). And once these drivers are identified it may be difficult to predict when they will be near each other.

Using systems biology to establish plausibility and insight has promise, but investigators should realize that other unknown models could mimic the same results. Therefore experimenters should consider how new experiments could distinguish among models (and perform those experiments before accepting any given model as valid). Using systems biology for reverse engineering could identify key components, whose interaction is the key to the system. However, a major peril is the likely variability of results when applying reverse engineering to gene expression arrays with thousands of genes. To ascertain variability, investigators should estimate networks in repeated random splits into training and test samples. Another promising aspect of systems biology is the prediction of cancer outcomes. However it is important to put these predictions in perspective, by comparing them with predictions based on models with clinical covariates.

Systems biology can advance hand-in-hand with empirical observations and confirmation. This process is enriched by collaborations across disciplines, including biostatisticians, theoretical and applied biologists, clinicians, and epidemiologists.

References

- Aijo T, Landesmaki H. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*. 2009; 25:2937–2944. [PubMed: 19706742]
- Ao P, Galas D, Hood L, Zhu X. Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. *Medial Hypotheses*. 2008; 70:678–684.
- An G, Mi Q, Dutta-Moscato J, Vodovotz Y. Agent-based models in translational systems biology. *Systems Biology and Medicine*. 2009; 1:159–171. [PubMed: 20835989]
- Baker SG. The central role of Receiver Operating Characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst*. 2003; 95:511–5. [PubMed: 12671018]
- Baker SG. Putting risk prediction in perspective: relative utility curves. *J Natl Cancer Inst*. 2009; 101:1538–1542. [PubMed: 19843888]
- Baker SG. Simple and flexible classification via Swirls-and-Ripples. *BMC Bioinformatics*. 2010; 11:452. [PubMed: 20825641]
- Baker SG, Kramer BS. Identifying genes that contribute most to good classification in microarrays. *BMC Bioinformatics*. 2006; 7:407. [PubMed: 16959042]
- Baker SG, Kramer BS. Paradoxes in carcinogenesis: New opportunities for research directions. *BMC Cancer*. 2007a; 7:151. 2007. [PubMed: 17683619]
- Baker SG, Kramer BS. Peirce, Youden, and Receiver Operating Characteristic Curves. *The American Statistician*. 2007b; 61:343–346.
- Baker SG, Soto AM, Sonnenschein C, Cappuccio A, Potter JD, Kramer BS. Plausibility of stromal initiation of epithelial cancers without a mutation in the epithelium: a computer simulation of morphostats. *BMC Cancer*. 2009; 9. [PubMed: 19134212]

- Chuang H, Lee E, Liu Y, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*. 2007; 3:140. [PubMed: 17940530]
- Fujita A, Patriota AG, Sato JR, Miyano S. The impact of measurement errors in the identification of regulatory networks. *BMC Bioinformatics*. 2009; 10:412. [PubMed: 20003382]
- Laubenbacher R, Hower V, Jarrah A, Torti SV, Shulaev V, Mendes P, Torti FM, Akman S. A systems biology view of cancer. *Biochimica et Biophysica Acta*. 2009; 1796:129–139. [PubMed: 19505535]
- Lemon G, King JR, Byrne HM, Jensen OE, Shakesheff KM. Mathematical modeling of engineered tissue growth using a multiphase porous flow mixture theory. *J. Math Biol*. 2006; 52:571–94. [PubMed: 16463188]
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. ARACNE. An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. 2007; 7(Suppl 1):57.
- Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*. 2005; 365:488–92. [PubMed: 15705458]
- Pittelkow YE, Wilson SR. Simpler evaluation of predictions and signature stability for gene expression data. *Journal of Biomedicine and Biotechnology*. 2009 2009; DOI:1155/2009/587405.
- Potter JD. Morphostats, morphogens, microarchitecture and malignancy. *Nat. Rev. Cancer*. 2007; 7:464–474. [PubMed: 17522715]
- Rosenfeld S. Characteristics of transcriptional activity in nonlinear dynamics of genetic regulatory networks. *Gene Regulation and Systems Biology*. 2009; 3:159–179. [PubMed: 20054406]
- Simon R. The use of genomics in clinical trial design. *Clin Cancer Res*. 2008; 14:5984–5994. [PubMed: 18829477]
- Smithers DJ. Cancer. An attack on cytologism. *The Lancet*. 1962; 279:493–499.
- Sonnenschein C, Soto AM. Theories of carcinogenesis: An emerging perspective. *Semin. Cancer Biol*. 2008; 18:372–377. [PubMed: 18472276]
- Soto AM, Sonnenschein C. The somatic mutation theory of cancer: growing problems with the paradigm. *BioEssays*. 2004; 26:1097–1107. [PubMed: 15382143]
- Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*. 2009; 27:199–204.
- Wang E, Lenferink A, O'Connor-McCourt MO. Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol Life Sci*. 2007; 64:1752–1762. [PubMed: 17415519]
- Werhli AV, Grzegorczyk M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*. 2006; 22:2523–2531. [PubMed: 16844710]

Figure 1.
Hypothetical ROC curve illustrating different slopes