# An Introduction to the Mathematical Modeling in the Study of Cancer Systems Biology

Abdallah K Alameddine[1], Frederick Conlin[2,3] and Brian Binnall[1]

[1]Division of Cardiac Surgery, Baystate Medical Center, Springfield, MA, USA. [2]Department of Anesthesiology, Baystate Medical Center, Springfield, MA, USA. [3]Division of Cardiac Surgery, University of Massachusetts Medical School, Worcester, MA, USA.

**ABSTRACT**

**BACKGROUND:** Frequently occurring in cancer are the aberrant alterations of regulatory onco-metabolites, various oncogenes/epigenetic stochasticity, and suppressor genes, as well as the deficient mismatch repair mechanism, chronic inflammation, or those deviations belonging to the other cancer characteristics. How these aberrations that evolve overtime determine the global phenotype of malignant tumors remains to be completely understood. Dynamic analysis may have potential to reveal the mechanism of carcinogenesis and can offer new therapeutic intervention.

**AIMS:** We introduce simplified mathematical tools to model serial quantitative data of cancer biomarkers. We also highlight an introductory overview of mathematical tools and models as they apply from the viewpoint of known cancer features.

**METHODS:** Mathematical modeling of potentially actionable genomic products and how they proceed overtime during tumorigenesis are explored. This report is intended to be instinctive without being overly technical.

**RESULTS:** To date, many mathematical models of the common features of cancer have been developed. However, the dynamic of integrated heterogeneous processes and their cross talks related to carcinogenesis remains to be resolved.

**CONCLUSIONS:** In cancer research, outlining mathematical modeling of experimentally obtained data snapshots of molecular species may provide insights into a better understanding of the multiple biochemical circuits. Recent discoveries have provided support for the existence of complex cancer progression in dynamics that span from a simple 1-dimensional deterministic system to a stochastic (ie, probabilistic) or to an oscillatory and multistable networks. Further research in mathematical modeling of cancer progression, based on the evolving molecular kinetics (time series), could inform a specific and a predictive behavior about the global systems biology of vulnerable tumor cells in their earlier stages of oncogenesis. On this footing, new preventive measures and anticancer therapy could then be constructed.

**KEYWORDS:** cancer biology, mathematical modeling, time series snapshots, biomarkers, systems biology

## Introduction

The pathogenesis of malignant transformation that leads to an inexorably progressing disease is intricately connected to the abnormal integration of several biological networks.[1–3] The objective analysis of the holistic approach in cancer research is to obtain a global quantitative description of all key interactions within the cell to ultimately predict how and why cells function the way they do in response to stressors such as hypoxia, DNA damage, retinoblastoma protein degradation, and oncogene activation or in response to external factors such as nutrient or oxygen deprivation.[4]

As we came to realize that the initial stage of tumor development follows a complex and multistep process that proceeds sequentially until a full-blown malignant phenotype is achieved and maintained, this begs the questions "what to do with and how to tie together this expanded type of data"?[5–7]

Fortunately, most of the tumors have a limited diversity in the key genes of cancer development and progression. Furthermore, functional consequences of cancer tissue heterogeneity may be minor.[8]

The aim of this report is a modest interpretive synopsis pertaining on how to model the dynamic behavior of interacting biological species in cancer.

## Materials and Methods

We begin this review with some mathematical definitions. We then focus on mathematical tools to model only sequential quantitative data set measurements overtime of those biochemical elements to systematically study the determinants of early cancer. This may prove useful in outlining new or emergent therapeutic prospects.[7,9]

### Important cancer biomarkers to consider when modeling parameters of interest

There are many possible underlying dysregulated signaling pathways in cancer.[2,3]

However, most of the components fall into the following broad categories that warrant consideration. There are those related to ligands/receptors and oncogenes such as PI3k/AKT,

TGF, MAPK, P53, RB, CDK, CYCLIN, TKR, NF-κB, miRNA, mTOR complex, P16, TERT, NOTCH, WNT, PTEN, and HEDGEHOG.[2] There are those related to the genes determining pluripotency such as c-MYC, OCT4, SOX2, NANOG, and KLF4. Expressions of those relevant molecular "populations" in these pathways are themselves under the control of multiple feedback loops and under tight homeostatic control and cross talk.[3,7]

Finally, there are those related to other core circuit systems and comprise the redox state such as cytosolic $NADH/NAD^+$ ratio, the mitochondrial electromechanical potential and mitochondrial transport chain, the energy charge inside the cell ($ATP+1/2\ ADP/\ ATP+ADP+AMP$ ratio), and the reactive oxygen species (ROS) scavenger $NADPH/NADP^+$ ratio.[3,10]

In this report, interaction with environmental factors and other as yet unknown processes will not be discussed here.[11]

## Results

### *Facing complex biological data in cancer research*

A mathematical model relates the dependent variables (such as a population growth or a particular molecular concentration growth) by means of a mathematical equation demonstrating the output cell response for a given input. Through molecular growth, we refer to a *time series* of concentration growth overtime of proteins, messenger RNAs (mRNAs), or other biomolecules all of which were cited above. Moreover, the model's objective is to quantify the system's behavior by the parameters that affect the process dynamics.

*Steady state or equilibrium.* Actual biological systems do not always exist in equilibrium because they contain many negative feedback loops by which the system's action is based on some previous state and not on the current state. Feedbacks are necessary so that the system may assume a resting and an active state. Because of the consequent time-delayed effects, this situation is likely to cause the system to miss or overshoot its goal either before the equilibrium target or before the "carrying capacity" of the system is reached (see infra). Studying equilibrium of a biologic process is important not because equilibrium reflects reality but because it is an important reference point for analyzing the final behavior of the model. The pure behavior of the model is then obtained by "shocking" the system out of equilibrium and knowing how the system responds to sudden input using a "step function." Equilibrium is stable if it does not change following this small but swift disturbance. A step function is a process that, when forced on the input, the output changes from zero and achieves a higher steady state in an instant. A steady-state level of any biomolecule is important for its optimal function. Computer programs can be set up to simulate the step function.

In mathematics, a function is a pairing of any set of inputs observed by changing the independent variables with their corresponding output responses. Functions can be linear or nonlinear, continuous or discrete. A function is continuous when its graph is a single unbroken curve. Polynomial functions, trigonometric, exponentials, logarithmic, and root functions are all continuous functions at every number in their domains.[12]

For an exponential input, of first-order or nth-order polynomial, the response will also be an exponential almost every time with the exception when there is resonance or near resonance between input and output frequencies.[13]

A graph is the best way to represent a function in which one element depends on another. Power functions and biological oscillations are typically nonlinear.

Cancer biology worlds are also nonlinear and are often discontinuous. In a discontinuous function, the rate of change at any point cannot be determined no matter how much we zoom in on that particular point. Linear systems exhibit growth, decay, equilibrium, or oscillations. A step function that is generated by a switch-like interaction can also be used to approximate the discontinuous function by a continuous one.[14]

*Critical or fixed points.* Fixed or critical points represent equilibria or steady states. Fixed points can be stable or unstable and only stable fixed point represents a stable equilibrium. It is also fundamental that one determines what is happening near the neighborhood of the critical points to estimate robustness as well as finding the limits of a given system's parameters. The critical points are found at the intersection of the graph, and either the 45° straight diagonal line of zero population growth or the identity coordinate $y$ (output) = $x$ (input), where the function's rate of change is 0 corresponding to constant operating conditions.[12] In other words, the diagonal line is where inputs get converted to outputs and vice versa.

Evaluating what happens near the fixed points is equivalent to a linearization technique at those points to determine the stability. If the system that begins close to the equilibrium approaches the steady state converges in a finite time reaching a fixed point, it is then stable; if it diverges from the fixed points and it grows indefinitely as time increases, it is then unstable, meaning it "approaches" infinity.[12,15] Cancer is considered to be an unstable process and by definition is neither in a steady state nor in equilibrium. Any dynamic system can settle down to equilibrium, keep repeating in periods, or become chaotic.[16] In contrast to oscillations, chaos never settles down to a periodic state and depends totally on the initial conditions.

For nonlinear dynamics in discrete time, it is simple to check for stability by linearization of the model locally, ie, when the time interval becomes shorter or using nonlinear ordinary differential equation (ODE). After linearizing the function, its transform is then computed (to eliminate dependence on time) and the result is just an easier algebraic function of the output. The output is then plotted using a graphing calculator. However, when we are dealing with too many interdependent nonlinear partial derivative equations, the disadvantage of linearization is that it becomes mathematically intractable and computationally prohibitive to implement.[17]

*Carrying capacity.* In nonlinear systems, when expressing the time evolution of molecular species of interest, the *logistic model* is used to predict the maximum growth limit in the long run. The simplest example of nonlinear population growth is the logistic equation or the sigmoidal stretched out "s" curve similar to the Hill function. Here, the molecular growth moves slowly, and then rapidly, before slowing down and tapering off through a negative feedback interaction. A critical stage in the evaluation of this function is to establish the upper value or the final steady state of the population size. The tapering off upper value is called the population's "carrying capacity."

When the molecular population is above the carrying capacity, it decreases, and when it is below the carrying capacity, it increases. In the logistic molecular growth model, the fixed points—here it means the final carrying capacity or the dependent variable—may switch stability as the parameter of growth rate (the independent variable) varies depending on the changes of external feedback factors. A more practical method to solve this type of growth capacity that is drifting in values is nondimensionalization of parameters. This is based on the fact that when parameters become dimensionless, the analysis is simplified because the parameters are reduced to either unity or less than unity. Nondimensionalization is obtained by dividing each parameter's value by the average of the total sum values or by their steady-state parameters value.

For future predictions of the maximum growth, applying the discrete logistic model is preferable to continuous exponential functions when there might be periodicity, sequences in the process, or when molecular successive generations do not overlap as time passes. When time is thus inherently discrete and consequently the carrying capacity becomes a drifting parameter, a more reliable model called the *logistic difference equation* should be considered.[14,15,17–19] The difference equations can easily be simulated on digital computers. Thus, the logistic function treats time as continuous, whereas the difference equation looks at discrete time steps.

*Multistability.* The presence of positive feedback loop can induce multiple steady states. Example of 2 possible equilibria or bistability is when considering a system composed of 2 genes that express transcription factors regulating each other, but in a negative manner.[20] A bistable system that can be constantly switched between 2 steady states following a transit input perturbation is called a toggle switch.[20] Bistability of regulatory circuits occurs when the rate of positively autoregulated gene product, by means of positive feedback loops, is strong compared with its degradation rate. MAPK is an example of a bistable signal in which the feedback loop enhances the robustness of 2 stable output states in response to a stimulus.[21] Moreover, on interconnection, such as in gene circuit, where the behavior of downstream system significantly affects the behavior of the upstream system, multistability may occur. The generally accepted reason for this behavior is that the activated

state ignores random small changes of signals (noise). If there are environmental heavy changes, some phenotypes will keep on ensuring survival. Furthermore, bistability generally widens the distribution of protein concentration per cell in which the distribution of proteins becomes bimodal.[22] Multistability includes an additional intermediate state called the hybrid state.

*Multistable nonlinear dynamics.* This is a system with multiple state equilibria in which a small variation leads to diverse steady states. Multistable systems cannot be described by a deterministic continuous equation but by a discrete stochastic modeling that evolves according to discrete jump Markov process. This leads to a probabilistic description of the system dynamics, ie, the likelihood of states and output variables is random and whose properties (mean, standard deviation) are governed by a probability density function (PDF).[17]

In a nonlinear dynamic system, there are interactions between its elements—part of the output has a feedback loop into the input. Unlike a time-independent Markovian process, the nonlinear dynamic is recursive and is time-dependent on previous events.[23] Nonlinear dynamic inputs of numerical data can be transformed to a reduced-order model which is derived using dynamic mode decomposition (DMD) algorithm. Dynamic mode decomposition extracts from flow of data only dynamically relevant or dominant features that contribute to the overall evolution of sequence.[24,25] Dynamic mode decomposition is a dimensionality reduction technique and is data driven where only snapshots and a set collection of time series of state measurements are needed.[24]

*Stochastic systems.* In biology, a stochastic system is a dynamic biological system that displays a noisy behavior possibly because of random fluctuation of all or parts of its components,[26-28] for example, the stochastic noise related to production and degradation of RNA molecules and the stochasticity from association and from dissociation of the transcription factors in the gene regulatory networks. Because of the stochastic noise, repeated experimental data do lead to different results. Because the stochastic system is a Markov process, several update functions need to be defined to formulate the system.[28,29] Based on the idea that a Markov process is aperiodic and irreducible, its probability distribution will end up in a stationary state in the long run. Its probability distribution can be approximated in a computationally favored way using a likelihood-based Markov chain Monte Carlo (MCMC) method. This is a repeated random sampling method from a uniform probability distribution (0, 1). The stable stationary state of Markov process allows the identification of RNA molecular PDF and by which deterministic approximation of the process is valid.[29]

Stochastic systems may have a deterministic linear component that is due either to an external event such as environmental

changes or to internal fluctuations inherent to random binding and unbinding of molecular elements. One important indication of stochastic behavior is bimodality or multiequilibria.[27,29]

As alluded to above, ODE cannot be applied to solve stochastic process which is usually discrete. Stochastic process can be otherwise described only using stochastic calculus in which the random function is approximated by dividing it into infinite numbers of small time steps which are then linear.

The uncertainty or the stochastic effect perturbing the system is generally assumed to have a Gaussian or normal PDF of mean zero and variance $\sigma 2$, similar to continuous version of a random walk or to the dynamics of diffusion-based phenomena. Solving the random function, ie, stochastic part is available in MATLAB (Mathworks.com) which provides a random number generated from a normal PDF (0, 1), ie, of mean 0 and variance of 1. After many runs simulating the stochastic process numerous times, the average of the runs will converge to the deterministic solution.

For the time evolution of a large *ensemble* of stochastic systems and averaging out their stochastic effects, a "master equation" is applied.[27] This equation, which is a system of ODEs, expresses the dynamics of a *population* fraction in a particular state that jumps to other state. Because the probability of the next transition depends on the current state only, the master equation is considered a continuous-time discrete-space Markov chain.

In the simplest cases, the master equation is analytically solvable. For larger models, more computationally efficient simulations of growth have been developed. The MCMC types of simulations are used that produce a random walk through the possible states of the system.

A stochastic simulation algorithm is also available to describe each particular trajectory of *individual* stochastic element separately.[30] The reason for stochasticity is due to the presence of many independent parts interacting with feedback control elements, cooperatively or competitively that create robustness, bistability, or oscillating runaway behavior.[28,31–33]

*Robustness.* Because there is often redundancy in biological systems, any small alteration of system's variables even when part of particular pathways is removed, the system can still maintain its performance and ignores these disturbances. This resiliency to large perturbation is interpreted as robustness,[34–36] which is closely related to the notion of Claude Bernard's homeostasis. To test the robustness of findings, sensitivity and specificity analyses should be conducted. In cancer, the parameters are not evolutionary fine-tuned; therefore, malignant cells maintain their robustness by multiple synergistic interactions as we will read later.

*Hill function.* Based on the law of mass action of chemical kinetics, the reaction rate is proportional to the product of the reactants' concentration. The kinetics of the equation is described by parameters called rate constants. In this situation, the metabolic rate is constant at high substrate concentration, thereby equivalent to a steady state. However, often enough enzymes have more than one binding site and their affinity changes. Positive cooperativity among macromolecules in enzymatic reaction at a steady state can be described by a sigmoid Hill function, which is a wide-spread function in modeling nonlinear responses, and in particular the regulatory pathways. Nonlinear relation can be log-transformed into a linear function.

Gene transcription networks are usually viewed as interconnected transcriptional components. An example for the use of a Hill function is relating the input of the production rate of a gene to the transcription factor concentration. While the Hill function makes the response faster, interconnectivity has an opposite effect.

*Oscillating functions.* Oscillations are common dynamic behaviors because they determine the most needed information of "delays" in the biological world of networks.[36,37]

Examples of oscillations in time are the TNF gene expression, oscillations in NF-κB (nuclear factor κB), and p53 signaling, and hormone secretion.[38] These systems are inherently nonlinear and exhibit varying behavior over time.

Negative feedback loops affect an oscillatory behavior. Positive feedback loops produce robustness and facilitate bistability.[35,36]

Oscillation means that if the system is perturbed slightly, it always returns to its periodic orbit. Cardiac rhythm, circadian rhythm, hormone secretions, and certain biochemical reactions oscillate spontaneously. Another example is glycolysis, a biochemical metabolic process that proceeds in oscillatory pattern.

If there is stochastic transcription of mRNA, stochastic synthesis of proteins, and in addition there is stochastic degradation of the same elements, the global behavior leads to a nonlinear dynamical system. Such example involves parameters that govern the degradation rates of both proteins and mRNA. To estimate the stable stationary-state solution and to solve these states, the response is transformed. At any fixed time, any periodic function with interacting parameters can be transformed into a sum of harmonics or frequencies with amplitudes or modes depending on the time at which the snapshots were taken. Most of the higher order exponential terms are ignored. Thus, the random distribution of inputs will end up with a simple system and lower dimension ODEs with a corresponding real solution.[15,26]

*Nonequilibrium systems.* A random type of motion, termed Brownian motion of a small macroscopic particle due to collisions with other molecules of the fluid when immersed in a liquid, may be applied to approximate the dynamics of nonequilibrium systems. Based on the central limit theorem, the

equilibrium PDFs of the fluctuating stochastic noise or "error" obey a Gaussian PDF.[15]

As alluded to earlier and because this process is discrete, it cannot be linearized and ODE cannot be applied. A statistical description by probabilistic manner is used instead. A computer-based MCMC simulation then determines the time evolution of the drift's PDFs from the deterministic trajectory, thus removing the uncertainty of the probabilistic component of the random variable's movement.[4,18]

*Discerning chaos and chaotic system's prediction.* In a simple linear or periodic system, when taken in isolation, its future behavior is well determined regardless of the initial starting points. In contrast, when that system is acted on by multiple dynamic and nondeterministic components ($\geq$3 dimensions), different long-term trajectories will occur.[16] In this higher dimension, the motion becomes chaotic. Overall, chaos is not equivalent to randomness because early on it is predictable (deterministic) but within a "time horizon" after which the system becomes effectively unpredictable.[16] Examples of chaotic systems are the movement of the galaxies, the weather, and in biological context, tumor cells; each system has its own time horizon that spans from days (for weather prediction) to millions of years (for galaxies).

After a transient time, the system settles into an irregular self-similar oscillation that persists without ever repeating and becomes aperiodic.[16] This view was labeled as Panta Rhei—meaning everything flows or "all entities move and nothing remains still"—by Heraclitus more than 2500 years ago. We now call this condition "fractal."

The distance or the separation between the trajectories grows exponentially fast, on average. Their rate of separation is the property of the system. Their distances follow a simple rule and can be quantified as a function of time. Because it is exponential, the logarithm of the distance is approximated by a trending line that is growing in time. The system can also compound or be damped nonlinearly overtime and self-similar without ever actually repeating.[16] For very complicated nonlinear dynamics with 1 degree of freedom and in which time is discrete, the difference equation can be applied at least early on in the deterministic transient period of the system.[16] Chaos then can have a simplistic mathematical model to represent it, and its dynamics is then defined by a recursive iterated steps. The population level at any given time step depends on the growth rate parameter and the previous time step's population level.[39-41] Thus, iterating the process many times by changing the growth rate parameter allows us to visualize the global behavior at a glance in a meaningful way.

*You have gathered the experimental biological data in a series of snapshots, now what?*

In this section, we provide the broad overview for modeling time-evolution analysis that can be applied to describe the behavior of a complex process such as tumorigenesis. Before analyzing the temporal sequential data points, the data should initially be normalized by either log-transforming or by dividing each value by their mean or by their steady-state value across total time interval.

For the analysis of high-dimensional time evolution data analysis, the available steps are given in the following sections.

Intuitively, visualize the data and make a rough sketch of scatter plot of each time point in the data set.[12,14] The next step is to graph the time series of the independent values to the dependent ones, and then check what the function looks like using computer programs. The packages MATLAB, Pylab, and Mathematica have such programs. Is the graph linear or polynomial (quadratic, cubic, or quartic, or higher order)? Is it periodic trigonometric, or is it power or exponential function? The exponential function grows much faster than the power function and is of interest because it occurs very frequently in mathematical models of nature. Finally, is there a need to reflect, shift, stretch, or compress the function to obtain the desired graph?

*Ordinary least square regression.* We build and graph the ordinary least square (OLS) that allows us to deduce the model with the use of a graphing computer, and the best fitting polynomial is chosen. Now the question is, "How tight is the fit?" To get the perfect fit, we use the coefficient of determination or $R2$ value for test data. This coefficient is dimensionless and scale independent with a range from 0 (worst) to 1 (best). The higher range of this coefficient means that the model fits tightly and that it explains most the variability in the estimates. So we select a model that has the highest $R^2$ value.

*Bayesian computation of large data.* This approach seeks to infer the parameters from the available observations. Bayesian classification can be used to classify the function (output) of variable attributes, numeric or categorical, based on the highest computer probability (likelihood) of the class that explains the given observations. The maximum likelihood is then chosen that explains the single best parameter which maximizes the observed data. Because writing the posterior distribution analytically becomes intractable in these cases, it will be necessary to use computational approaches, such as MCMC techniques which will be outlined later.[42,43] For instance, it is an approach that is suitable to address differential gene expression.[23] In the inferential statistics, the biggest impediment is that the prior or the apriori likelihood of a class (ie, without knowing the current attributes) remains unidentified. The other hindrance is that Bayesian classifier cannot differentiate classes if the attributes are correlated (ie, multicollinearity) with similar distribution parameters' PDFs.

*Solving the problem of multicollinearity.* When we are dealing with too many variables, which is often the case in cancer research, multicollinearity may be a problem. Multicollinearity means that a large set of potential covariates exist, and that

those variables are highly correlated with high variance between samples and with completely different values of the regression coefficients. In an ideal data set, we need high variations between the elements, meaning the diagonals of the data regression matrix, but we want small covariations, meaning all off-diagonal entries of the matrix. In that situation, neither OLS nor *t* test should be used.

What do we do if there is a problem with multicollinearity? The standard approach is to eliminate the variables that are closely correlated with the other variables and do not include them in the model.

*Ridge regression* is the preferred method, however, and more importantly, it is only used as diagnostic for multicollinearity. Here, we inflate the variances of the diagonals' matrix by a small "bias factor," less than 1 to match the covariance. If the coefficients of the independent variables change in magnitude or if they flip in sign when compared with those in the original OLS, then we chose the smallest amount of bias factor that keeps the slope estimates relatively stable.

The second method is using the principal component analysis (PCA) in which each sample is displayed and graphed. The distances (variability) between data points are represented on the coordinates in this graph. The largest distances provide information about the network's dominant properties. Unfortunately, these variabilities are completely decorrelated from each other, making this method too difficult to interpret, and the correlation with the original data becomes meaningless. Of note, in light of the fact that signaling networks are dynamic and nonlinear due to multiple feedback loops and that some are oscillatory, PCA-based methods cannot be applied in these sinusoidal mode behaviors.

The third method is "Centering." This is done by rescaling, ie, by subtracting each dependent variable or attribute from the mean. Its disadvantage is that it applies only when we have very few interaction terms.

Finally, if the number of relevant variables surpasses the number of available observations, then the least absolute shrinkage and selection operator (LASSO) regression should be used to achieve sparsity.[44] The LASSO further reduces the coefficients of regression by eliminating the ones with zero values and keeping the nonzero ones thus achieving sparsity and therefore there will be fewer features to keep in the model.

*In silico validation.* All models need to be validated before applying them for long-term prediction.[43,45]

Due to the presence of noise (or error) in estimating the parameters, we cross-validate to determine whether the model can be applied to an independent data set in the real world. This is done by splitting the data into a training data set and a testing data set. The next step is to choose the set with the lowest error. Then, the simplest model possible not only accounts for the data at hand but also predicts new data effectively. Other ideas based on cross-validation include use of resam-

pling to directly estimate the prediction error and leave-one-out cross-validation when the data set is not too large.

*MCMC technique and simulation.* The MCMC is actually similar to a random walk that models each state at each time point as a linear process that proceeds in steps. When the next step occurs, the process can be in the same state or moves to another state. All movements between states are represented by a probability and the sum of each state's probability must add up to one. This model will find the probability of being in the final steady state many steps into the process.[4] The time that each process spent in each state before it jumps to the next state has an exponentially distributed waiting time or a Poisson distribution, for example, reading off the DNA template by RNA polymerase.

The results from the analysis that is based on only few observations can be highly subjective because the analysis could have missed other possibilities. For the inquiry to be more reliable or more accurate to simulate reality, we must figure out the PDFs of each unknown feature. A more precise analysis is done by computer simulation in which the computer generates, for each unknown factor, a random number that is between 0 and 1. The simulation applies each random number to each unknown factor's PDF and finds the value corresponding to the random number PDF. The process is repeated to predict the final distribution. Running numerous iterations including random number sampling to simulate a noise model can be applied. This method leaves us better informed about the final decision because it is based on all possible outcomes and not only restricted to a very few.

## Applications of Basic Tools in Computational Modeling of Tumorigenesis

*Quantitative modeling of complex gene regulation in eukaryote from proteomic data*

A unifying model of carcinogenesis implicates the induction of genetic instability (somatic mutations), inflammation, and the release of microenvironmental forces. Specifically, these are the regions of intermittent and chronic hypoxia, ROS, and the acidotic milieu generated by active glycolysis. The synergistic action of those processes feed backs DNA genotoxicity and increases the accumulation of further mutations, inflammation, immune suppression, and tumor growth. Furthermore, cell death response is inhibited, and more hypoxia-induced glycolysis and acidosis accumulate.[46] The dynamics of somatic mutations can be modeled as the main drivers of carcinogenesis.[47] Once mutation rate of a particular gene has been known, the alteration likelihood of cardinal signaling pathway can then be defined.[48] These signaling pathways are involved in balancing energy requirement, survival, cell cycle progression through the restriction point, and the biomass formation.

The transition between on-and-off gene activities is considered a stochastic multistep process. This switch depends on a

set of sequentially assembled factors for histone modifications. Its description includes a specific Markovian chain. Assuming that the lifetime of mRNAs is short compared with that of proteins, it can be integrated out; then, the burst size is geometrically distributed and the initiation waiting time of the on-state is exponential. As observed experimentally available from single cell data, proteins are produced in random bursts resulting in a gamma-distributed steady-state protein concentration and exponentially distributed number of protein molecules per burst.

Gamma distribution in this context depends on 2 parameters: the burst size and the mean number of proteins produced per burst. Analytical solution to stochasticity in gene expression can then be linked to the steady-state distribution of a given protein concentration in a cell population.[26]

## Computational analysis of nonlinear oncogenic and inflammatory signals

To identify the key component of a biological system, a basic numerical method is used which approximates the curve function with a series of slopes (rate of change of the reaction) to monitor a species concentration over time.[21] There are other more advanced numerical integration of complex reactions with Michaelis-Menten kinetics and comprising multiple biochemical species that concurrently solves all ODEs of the system's attributes.[49]

The binding of ligand to receptors such as the insulinlike growth factor or to the epidermal growth factor allows the activation of various signaling pathways such as the PI3K-AKT, EGF, and MAPK/ERK pathway. MAPK maintains energy balance at both the cellular level and whole body level. It monitors the cellular ratios of AMP/ATP and ADP/ATP.

Mammalian target of rapamycin complex 1 (mTORC1) is a serine/threonine kinase that promotes ribosomal protein synthesis and controls cellular growth by monitoring nutrient availability, cellular energy status, oxygen levels, and mitogenic signals. Activation of mTORC1 is achieved through phosphatidylinositol 3-kinase and AKT. Aberrant activation of the mTORC1 pathway has been implicated in a variety of cancers. The E2F-RB pathway controls the G1-to-S-phase transition through the restriction point by the timely activation of genes involved in DNA synthesis and cell cycle control. Experimental evidence, based on regression-type analyses, show that E2F-RB pathway is a nonlinear bistable switch.[50] The switch is generated by 2 positive feedback loops. As we learned this notion before, bimodularity—meaning high and low different concentrations in different cells—is produced by stochastic fluctuations of protein concentrations. Activation by growth factors stimulation relieves inhibition of the transcription factor E2F by RB and allows the migration of E2F through a cascade of events leading to the activation of multiple target genes and cellular proliferation. The most common genetically mutated

pathways in human malignancies are the E2F-RB pathway and the p53 tumor suppressor pathway.

Inflammation involves multiple signaling pathways. When applied to time series data in the blood, PCA can identify the critical drivers of inflammatory phenotype that have most influence on cancer progression. The transcription factor NF-κB is particularly important in the regulation of inflammation, apoptosis, and proliferation in cancer cells. Inducers of NF-κB include inflammatory cytokines, growth factors, lipopolysaccharide, oxidative stress, and viral products. Experimental and computer simulation implicated NF-κB in the control of TNF-induced apoptosis of the inflamed tissues.

Monte Carlo strategy, using sampling to generate random sets of parameter values for simulations, then recapitulates in silico key feature of the inflammatory mediators' dynamic. Computationally, this involves the concentration for each component and the kinetic rates for interactions to be determined. Data quantification of time series parameters is essential for prediction accuracy, and nonlinear ODE of biochemical pathways is then applied for that purpose.

Dynamic Bayesian networks are great frameworks to infer such large network interactions by determining how well the network can explain observed data. For pairwise nonsequential sampling of nonlinear systems (contained in the dependent and independent variables), DMD provides the best-fit linear approximation. In this method, computation can still identify the relationship between future and past values of time series data taken from multiple runs of an experiment to reduce the noise.[24,25,51] Dynamic mode decomposition is similar to the difference equation in which the recursive sampling for the data migrates the values of a parameter to the origin becoming sufficiently stationary. Dynamic mode decomposition is used to lower dimensions for dynamical systems with irregular time interval by collecting only few snapshots.

Ordinary least square and ridge regression both have drawbacks based on low predictive accuracy. Furthermore, they should be used only in linear dynamics. If moderate-sized effects in a large biologic data, Lasso is best.[44]

## Computational modeling of metabolic energetics and the immune response

In all living organisms, metabolites are derived from only 12 precursor metabolites such as glucose-6-P, ribose-5-P, fructose-6-P, erythrose-4-P, glyceraldehyde-3-P, oxaloacetate, 3-P-glycerate, P-enolpyruvate, 2-oxyglutarate, pyruvate, succinyl-CoA, and acetyl-CoA. Their biosynthesis requires the use of 3 pairs of cofactors such as ATP/ADP, NADH/NAD, and NADPH/NADP. In addition, there are approximately 50 building blocks which need to be included in any mathematical model chosen.[52]

The enzyme pyruvate kinase (PMK) isoform 2 catalyzes the crucial step in glycolysis, thus facilitating the switch from

oxidative phosphorylation to glucose metabolism and generating ATPs in cancer cells (Warburg effect). Nuclear translocation of PMK2 reprogramming is both activated by and activates hypoxia-inducible factor 1–dependent transcription. Both phosphofructokinase and PMK2 have been shown to generate glucose metabolic transition mediated by lactate and AKT.

Certain enzymes play critical roles in the metabolism of cancer. ALDH3B1 acts against oxidative stress derived from lipid peroxidation. IDO1 is involved in tryptophan metabolism and immune suppression.[3]

The citric acid cycle is catalyzed by 8 enzymes of which α-ketoglutarate dehydrogenase has the lowest activity and thus is considered the rate-limiting enzyme in the cycle. Computational modeling of the mitochondrial respiratory and energy metabolism encompasses integrating a series of ODEs for each state variable with respect to time. Model simulation is run till it reaches a steady state. The basic components included in the model are the reactions at electron and substrates (ATP/ADP/AMP) transport system and the transporters/antiporters of $K^+/H^+$ cations with their own parameter values.[53]

Flux balance analysis is also available for dynamic simulation. Flux balance analysis is a mathematical model that uses the flow of metabolites through a stoichiometric metabolic network.[54]

The overcoming of immune incursion can be described by the predator-prey iterations (logistic function type model as we noted earlier) that is simulated by a population-based model between tumor and immune cells.[54] To accommodate their energy needs, immune cells use the glycolytic pathways. The immune check point ligand present on cancer cells can suppress the immune cells check point proteins and the glycolytic signaling pathways such as AKT and mTOR. This process impairs immune cells and leads to promoting cancer progression.[55]

### Study of biological oscillators

Sustained oscillations of interacting proteins or genes could be generated by negative feedback loops. Many biological oscillators including p53, cell cycle, RB-E2F, NF-κB, and cardiac and circadian rhythms also have positive feedback loops. The feedback loops transform the p53 responses from periodic pulses to a single sustained pulse. The p53 has major roles in genomic stability. In response to DNA damage, p53 can show a transient response leading to cell cycle arrest and DNA repair.[54] Pulsatile responses induce cell apoptosis, and sustained responses mainly trigger senescence. The additional loops have been shown, through computational studies and tested using Monte Carlo approach, to impart some performance advantages over simple negative feedback loops. Positive-plus-negative oscillations appear to be more robust and to have better adjustable frequency without compromising output. The

parameterized oscillator models were simulated over a wide range of the Hill function coefficients.[56]

### Epithelial–mesenchymal transition

In cancer, epithelial-mesenchymal transition (EMT) is a necessary switch from an epithelial to a mesenchymal phenotype that enhances metastatic cell dissemination. Different pathways, a variety of transcription factors and microRNAs (miRNAs), have been shown to activate EMT, including TGF-β, p53, NF-κB, WNT, Notch, EGF receptor (EGFR), and Hedgehog. Transcriptomic analysis has revealed that TGF-β is a master regulator of EMT.[57] The EGFR is most frequently altered in cancer through either overexpression or mutation. The ROS, high in hypoxic condition, leads to an enhanced production of TGF-β. Recent works focused more on computational systems biology models of EMT rather than on the kinetics of individual players in the network and have elucidated the underlying mechanisms of this transition.[54,58]

Epithelial-mesenchymal transition is regulated by intercalated systems of 2 positive and 2 negative feedback loops between miR-mediated reactions (including translational regulation, alternative splicing, and epigenetic modifiers) and EMT-inducing transcription factors such as ZEB1/2, SNAIL1/2, and TWIST1. Activation of EGFR induces high miR/low ZEB levels (epithelial phenotype), and TGF-β induces high ZEB/low miR levels (mesenchymal phenotype). The combination of these states can produce a multistable EMT phenotype.[54]

### Total integration of multi "omics" data

To acquire the expected value of a continuous variable, we integrate the nonzero values of the variable times the probability mass density (PMD). For a discrete and random variable, we use a sum instead of an integral. Getting PMD requires 2 steps: identifying the type of function by reviewing the data graph and then we look up the table of moment-generating function (MGF). The probability distribution of a random variable is uniquely determined by MGF.

Analyzing complex biological systems such as cancer comprises high-dimensional global sets. These data include stochastic genome transcriptions, epigenetic alterations, proteome, metabolome, and signaling molecules. These biological systems and their interactions maintain phenotype stability and robustness which are strongly chosen by evolution. Positive feedback enhances sensitivity, whereas negative feedback dampens diverse perturbations imposed by the environment ensuring stability for homeostasis.[59]

Both normal and cancerous cells use time integral of the error which is fed back into the system. The error is the difference between the actual output and the desired steady-state output. This type of control maintains the error as low as possible despite perturbations in the input.[60]

Because the examined number of independent variables is very high and the number of their associated samples is low, a proposed and easy model based on nonlinear regression can be implemented. When the system is first linearized and then its transfer function is determined, the other efficient algorithm uses an iterative approximation to estimate a solution for a system of linear equations. In a manner similar to autoregression algorithm, we use the efficient and preferred Gauss-Seidel method with an initial guess value to solve system of equations *iteratively* in which convergence is guaranteed if the matrix of coefficients is square and diagonally dominant. If no convergence, then we use stopping criteria in percentage. The other method is a finite difference method where the systems unknowns are solved *simultaneously* knowing the boundary constraints values.

Protein-protein interactions with a functional role can be selected based on their degree of centrality.[61] Network interconnections that are responsible for the behavior of the cell are dominated by a power-law degree distribution indicating that a few hubs hold the most redistribution points and enzymes are considered in their links.[62] The networks include protein-protein interactions, metabolic, signaling, and transcription-regulatory networks. Deciding which species are to be included in the evaluation of kinetic data is crucial. It should be based on both the previously accumulated experimental research in the literature and the theoretical intuitive considerations.

## Conclusions and Outlook

The intricately connected genetic somatic mutations/epigenetic alterations and the environmental cues differentially affect cancer initiation and clonal expansion making cancer an intriguing disease to decipher.[1–3,63] The extensive progress that has elucidated the complex biological functions necessary for cellular communications has led to improvement in our understanding related to cell cancer survival and proliferation.[3]

For example, to simulate temporal dynamics of molecular networks in cancer, kinetic modeling has revealed transient, oscillatory, sustained, or toggle-switch-like systems properties among cancer regulatory networks. In these situations, it is crucial to identify the nodes that control this dynamic circuitry. These different specific adaptation states and response dynamics depend greatly on the amplitude and duration of signals as well as the different feedback adjustments.

A high-throughput biological technology has resulted in a large amount of high-dimensional data that we are only now starting to investigate cancer as an entire system.[42,43]

Is mathematical modeling all that is needed to simulate the real world?[63] The answer is not conclusive because interaction with environmental factors and other as yet unknown processes remains not well understood. Furthermore, it remains unclear whether what we are modeling captures the essence of how cancer cells operate. The answer is becoming increasingly distinct, in the near future. With further refinement in mathematical modeling as outlined in this review and experimental validation, it may become even pure.

## Author Contributions

AKA and FC conceptualized the manuscript. AKA and FC helped in methodology of the data. AKA and FC investigated the manuscript. AKA, FC, and BB analysed the data. AKA and FC resources the data. AKA and FC wrote the original draft of the manuscript. AKA, FC, and BB contributed to the writing, review and editing of the manuscript. AKA and BB visualized the manuscript. AKA supervised the manuscript.

## REFERENCES

1. Bradner JE, Hnisz D, Young RA. Transcriptional addiction in cancer. *Cell*. 2017;168:629–643.
2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–674.
3. Alameddine AK, Conlin FT, Binnall BJ, Alameddine YA, Alameddine KO. How do cancer cells replenish their fuel supply? *Canc Rep*. 2018;1:e003.
4. Vera J, Wolkenhauer O. Mathematical tools in cancer signaling systems biology. In: Cesario A, Marcus FB, eds. *Cancer Systems Biology, Bioinformatics and Medicine. Research and Clinical Applications*. New York, NY: Springer; 2011:185–212.
5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339:1546–1558.
6. Aebersold R, Aufrey C, Baney E, et al. Report on EU-USA workshop: how systems biology can advance cancer research. *Mol Oncol*. 2008;3:9–17.
7. Del Vecchio D, Murray RM. Introductory concepts. In: Del Vecchio D, Murray RM, eds. *Biomolecular Feedback Systems*. Princeton, NJ: Princeton University Press; 2014:1–28.
8. Meric-Bernstam F, Frampton GM, Ferrer-Lozano J, Yelensk R. Concordance of genomic alterations between primary and recurrent breast cancer. *Mol Cancer Ther*. 2014;13:1382–1389.
9. Prentice RL. On the role and potential of high-dimensional biologic data in cancer research. In: Li X, Xu R, eds. *High-Dimensional Data Analysis in Cancer Research*. New York, NY: Springer; 2009:1–11.
10. Berg JM, Tymoczko JL, Gatto GJ Jr, Stryer L. Regulation of mitochondrial ATP synthase. In: Berg JM, Tymoczko JL, Gatto GJ Jr, Stryer L, eds. *Biochemistry*. 8th ed, Part II, chapter 18. New York, NY: W. H. Freeman & Company; 2002.
11. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nature Biotechnol*. 2010;28:245–248.
12. Stewart J. Functions models. In: Stewart J, ed. *Calculus Early Transcendentals*. 5th ed. Belmont, CA: Thomson Brooks/Cole; 2003:10–63.
13. Stewart J, Day T. Applications of derivatives. In: Stewart J, Day T, eds. *Biocalculus: Calculus for the Life Sciences*. Boston, MA: Cengage Learning; 2015:250–306.
14. Stewart J, Day T. Functions sequences. In: Stewart J, Day T, eds. *Biocalculus: Calculus for the Life Sciences*. Boston, MA: Cengage Learning; 2015:2–80.
15. Barillot E, Calzone L, Hupe P, Vert JP, Zinovyev A, eds. *Computational Systems Biology of Cancer*. Boca Raton, FL: Chapman & Hall/CRC; 2012.
16. Strogatz SH. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Reading, MA: Perseus Books; 1994:348–397.
17. Wiggins S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. New York, NY: Springer; 2003.
18. Harezlak J, Tchetgen E, Li X. Variable selection in regression—estimation, prediction, sparsity, inference. In: Li X, Xu R, eds. *High-Dimensional Data Analysis in Cancer Research*. New York, NY: Springer; 2009:13–58.
19. Wilkinson DJ. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet*. 2009;10:122–133.
20. Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*. 2000;403:339–342.
21. Orton RJ, Sturm OE, Vyshemirsky V, Calder M, Gilbert DR, Kolch W. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. *Biochemical J*. 2005;392:249–261.
22. Alon U, ed. *An Introduction to Systems Biology: Design Principles of Biological Circuits (Mathematical and Computational Biology)*. Boca Raton, FL: Chapman & Hall/CRC; 2007.
23. Jain S. Bayesian approaches: nonparametric Bayesian analysis of gene expression data. In: Li X, Xu R, eds. *High-Dimensional Data Analysis in Cancer Research*. New York, NY: Springer; 20087:1–20.
24. Proctor JL, Brunton SL, Kutz N. Dynamic mode decomposition with control. *SIAM J Appl Dyn Syst*. 2016;15:142–161.

25. Koopman BO. Hamiltonian systems and transformation in Hilbert space. *Proc Natl Acad Sci U S A*. 1931;17:315–318.
26. Friedman N, Cai L, Xie S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett*. 2006;97:168302.
27. Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys*. 1976;22:403–434.
28. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*. 2008;135:216–226.
29. Tierney L. Markov chains for exploring posterior distributions. *Ann Statist*. 1994;22:1701–1728.
30. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 1977;81:2340–2361.
31. Åström KJ, Murray RM. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press; 2008. http://www.cds.caltech.edu/~murray/amwiki/index.php/Second_Edition.
32. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A*. 2002;99:12794–12800.
33. Nelson DE, Ihekwaba AE, Elliott M, et al. Oscillations in NF-kappaB signaling control the dynamics of gene expression. *Science*. 2004;306:704–708.
34. Sun L, Yang G, Zaidi M, Iqbal J. TNF-induced gene expression oscillates in time. *Biochem Biophys Res Commun*. 2008;371:900–905.
35. Barkai N, Leibler S. Robustness in simple biochemical networks. *Nature*. 1997;387:913–917.
36. Tsai TY, Choi YS, Ma W, Pomerening JR, Tang C, Ferrell JE. Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science*. 2008;321:126–129.
37. Tonnelier A. Cyclic negative feedback systems: what is the chance of oscillation? *Bull Math Biol*. 2014;76:1155–1193.
38. Lahav G, Rosenfeld N, Sigal A, et al. Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nat Genet*. 2004;36:147–150.
39. May R. Simple mathematical models with very complicated dynamics. *Nature*. 1976;261:45967.
40. Boeing G. Visual analysis of nonlinear dynamical systems: chaos, fractals, self-similarity and the limits of prediction. *Systems*. 2016;4:37. doi:10.3390/systems4040037.
41. Lorenz EN. Deterministic nonperiodic flow. *J Atmos Sci*. 1963;20:130–141.
42. Liepe J, Kirk P, Filippi S, Toni T, Stumpf MPH. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc*. 2014;9: 439–456.
43. Klipp E, Krause F. Computational tools for systems biology. In: Cesario A, Marcus F, eds. *Cancer Systems Biology, Bioinformatics and Medicine Research and Clinical Applications*. New York, NY: Springer; 2011:213–243.
44. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Meth*. 1996;58:267–288.
45. Müller J, Kuttler C, eds. *Methods and Models in Mathematical Biology: Deterministic and Stochastic Approaches*. New York, NY: Springer; 2015.
46. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics unifies carcinogenesis and cancer therapy. *Nat Rev Cancer*. 2012;12:487–493.
47. Derbal Y. Perspective on the dynamics of cancer. *Theor Biol Med Model*. 2017;14:18.
48. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013;502:333–339.
49. Press WH, Taukolsky SA, Vetterling WT. *Numerical Recipes: The Art of Scientific Computing*. 3rd ed. Cambridge: Cambridge University Press; 2007:899–946.
50. Kolch W, Halasz M, Granovskaya M, Kholodenko BN. The dynamic control of signal transduction networks in cancer cells. *Nat. Rev. Cancer*. 2015;15:515–527.
51. Tu JH, Rowley CW, Luchtenburg DM, Brunton SL, Kutz JN. On dynamic mode decomposition: theory and applications. 2014;1:391–421. doi:10.3934/jcd.2014.1.391.
52. Nielsen J. Systems biology of metabolism. *Annu Rev Biochem*. 2017;86:245–275.
53. Daniel A, Beard DA. A biophysical model of the mitochondrial respiratory system and oxidative phosphorylation. *PLoS Comput Biol*. 2005;1:e36.
54. Magi S, Iwamoto K, Okada-Hatakeyama M. Current status of mathematical modeling of cancer—from the viewpoint of cancer hallmarks. *Curr Opin Syst Biol*. 2017;2:38–47.
55. Wong BW. Metabolism of stromal and immune cells in health and. 2014;511:167–176.
56. Tsai TY, Choi YS, Ma V, Pomerening JR, Tang C, Ferrell JE Jr. Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science*. 2008;321:126–129.
57. Aglianono A, Calvo A, Box C. The challenge of targeting cancer stem cells to halt metastasis. *Semin Cancer Biol*. 2017;44:25–42.
58. Jolly MK, Levine H. Computational systems biology of epithelial-hybrid-mesenchymal transitions. *Curr Opin Syst Biol*. 2017;3:1–6.
59. Becskei A, Serrano L. Engineering stability in gene networks by autoregulation. *Nature*. 2000;405:590–593.
60. Yi TM, Huang Y, Simon MI, Doyle J. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci U S A*. 2000;97:4649–4653.
61. Cava C, Bertoli G, Colaprico A, Olsen C, Bontempi G, Castiglioni I. Integration of multiple networks and pathways identifies cancer driver genes in pan-cancer analysis. *BMC Genomics*. 2018;19:25.
62. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5:101–113.
63. Greene JA, Loscalzo J. Putting the patient back together—social medicine, network medicine, and the limits of reductionism. *N Engl J Med*. 2017;377:2493–2499.