Review

Chong Yu and Jin Wang*

Data mining and mathematical models in cancer prognosis and prediction

https://doi.org/10.1515/mr-2021-0026
Received September 27, 2021; accepted December 29, 2021;
published online June 29, 2022

Abstract: Cancer is a fetal and complex disease. Individual differences of the same cancer type or the same patient at different stages of cancer development may require distinct treatments. Pathological differences are reflected in tissues, cells and gene levels etc. The interactions between the cancer cells and nearby microenvironments can also influence the cancer progression and metastasis. It is a huge challenge to understand all of these mechanistically and quantitatively. Researchers applied pattern recognition algorithms such as machine learning or data mining to predict cancer types or classifications. With the rapidly growing and available computing powers, researchers begin to integrate huge data sets, multi-dimensional data types and information. The cells are controlled by the gene expressions determined by the promoter sequences and transcription regulators. For example, the changes in the gene expression through these underlying mechanisms can modify cell progressing in the cell-cycle. Such molecular activities can be governed by the gene regulations through the underlying gene regulatory networks, which are essential for cancer study when the information and gene regulations are clear and available. In this review, we briefly introduce several machine learning methods of cancer prediction and classification which include Artificial Neural Networks (ANNs), Decision Trees (DTs), Support Vector Machine (SVM) and naive Bayes. Then we describe a few typical models for building up gene regulatory networks such as Correlation, Regression and Bayes methods based on available data. These methods can help on cancer diagnosis such as susceptibility, recurrence, survival etc. At last, we summarize and compare the modeling methods to analyze the development and progression of cancer through gene regulatory networks. These models can provide possible physical strategies to analyze cancer progression in a systematic and quantitative way.

Keywords: flux; gene regulatory network; landscape; machine learning; ordinary differential equations; stochastic differential equations.

Introduction

By 2018, about 9.6 million people worldwide had died of cancer, and cancer has become recognized as the second leading cause of human death [1]. Moreover, deterioration of the global ecosystem has contributed to a spurt in the cancer incidences. It is expected that by 2030, the number of cancers will reach 23.6 million [2], more than double of the current number. Cancer is a generic term to a kind of fetal and complex disease. In the majority of cancers, the disease affect different tissues of the body. The normal cells are transformed to cancer cells in a multistage process and finally turn into malignant cancer cells. These variations include genetic (genes) and epigenetic (physical and chemical carcinogens, and biological infections) changes.

There are approximately three billion base pairs in human DNA. About 20% of the human DNAs are used to encode the proteins, and the other 80% are used to encode the retrotransposons, transposons and pseudogenes [3, 4]. Meanwhile, DNAs also encode many types of microRNAs. With the use of microarray and RNA-sequencing data, researchers began to monitor transcriptomes on a genome-wide scale, which has dramatically expedited comprehensive understanding of the gene expression profiles. The enormous amounts of high throughput data have been collected and available waiting for analysis and processing. Accurate prediction and classification are one of the major important and challenging issues for medical professionals [5]. At the same time, computer technology is

***Corresponding author: Jin Wang,** Department of Chemistry and of Physics and Astronomy, State University of New York at Stony Brook, Stony Brook, NY 11794-3400, USA,
E-mail: jin.wang.1@stonybrook.edu
Chong Yu, State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun, Jilin, China; and Department of Statistics, JiLin University of Finance and Economics, Changchun, Jilin Province, China. https://orcid.org/0000-0003-1289-8578

also developing rapidly which provides strong support for researchers to solve the problem of large scale data computing. These enable researchers to utilize automation and computer-based modeling for efficient processing of large data sets. Computational approaches have been developed to solve these kinds of complex calculation problems and help to get more information about cancer. Data mining and the machine learning (ML) methods become an essential tool for researchers [6, 7]. These techniques can be used for diagnosis and prognosis of different cancer diseases from the complex datasets or predicting cancer susceptibility, recurrence, survival etc. in future [8–10].

However, it is generally believed that analyzing a large amount of datasets does not necessarily provide comprehensive understanding of a given gene in a biological process. Many crucial biological processes and molecular pathways are based on the interactions and regulations among genes. The genes in cellular environment regulate each other and the gene expressions also reflect the result of regulation. In order to understand the cellular function, it is necessary to study the regulations of genes from a holistic perspective instead of from individual manner. Developing gene regulation models such as gene regulatory networks [11–13] can facilitate the quantitative understanding of these comprehension problem. Gene regulatory network can be inferred from the gene expression data such as RNA-seq [14], microRNAs data (miRNAs) such as microRNA (miRNA)-seq [15], protein-DNA interactomic data such as ChIP-seq [16] and so on. These data can serve as the raw input data to provide information on gene regulations and functions at the gene expression level. According to the gene expression information, gene-gene interactions can be inferred among different genes. Gene regulatory networks often have the information of the regulatory motifs, expression profiles and interactions between regulatory genes [17]. The importance of gene regulatory networks has become increasingly significant for all biological systems [18]. Furthermore, gene regulatory network can also serve as a working model to help the researchers to propose some hypotheses and assist in experimental design [19].

How to infer the actual gene regulatory network with these information and resources becomes an endless discussion. That is because different models have different characteristics and are suitable for different data types. In this review, we summarized several common machine learning methods such as Artificial Neural Networks (ANNs), Decision Trees (DTs), Support Vector Machine (SVM) and naive Bayes. We also introduced a few models which are popular in establishing the gene regulatory networks. These models are Correlation, Regression and Bayesian (native and dynamic) methods. Then we provide some examples to compare the analysis and accuracy. To improve the accuracy of the gene regulatory network construction some prior information has been added into certain of the methods.

Ordinary differential equations (ODEs) are mathematical methods for the descriptions of non-linear dynamical evolution of the system. Stochastic differential equations (SDEs) are the ODEs with added noise terms which describe the stochastic fluctuations. More specifically, SDEs also is often called 'Langevin' equations which can be used to describe dynamical systems [20, 21]. SDEs can be regarded as an extension of the dynamic system theory to the noisy or fluctuating regime. This is an important generalization, as real systems cannot be completely isolated from their environments, so they are always influenced by external sources. ODEs and SDEs are widely applicable on molecular dynamics. In this review, we also introduces analytical methods (ODE and SDE) based on gene regulatory networks to reveal the progress and development of cancer. In particular, we will introduce a landscape and flux theory to globally describe to dynamics of the understanding gene regulatory net work and apply the theory to cancer. Both data mining and mathematical models can help us in cancer prediction and diagnosis which can provide necessary information on cancer treatments.

Machine learning methods in cancer study

Machine learning methods

Machine learning (ML) is widely used for predicting patterns in a generalizable way [22]. ML exploits the computational methods to learn information directly from the historical data or experiences. Machine learning is usually divided into two main types: (i) supervised learning and (ii) unsupervised learning. In supervised learning, a labeled set of input data is used for the training and map to the desired output. After many iterations of this training, the model, when it receives a input, can give an output based on the experience it has learned. In contrast, the unsupervised learning methods provided no labeled input and there is no output during the learning process. For example, clustering is a typical unsupervised method. We need to artificially specify the number of clusters. Then the algorithm tries to put the data in different clusters in order to describe the data characteristics. During the process, each new sample can be put into the identified clusters which has the similar characteristics.

Once we built or validated a model, then it can be used for classification and regression or to check which parts of the data are relevant. Usually, there are four steps to apply machine learning technology to data sets. (a) Put forward a specific mathematical problem suitable for data statistical model. (b) Determine which specific model in the model class is most suitable for the data (this usually involves the numerical optimization function of some objectives to generate a set of fixed parameters to identify the specific model in the model class) (c) Use test sets or cross validation to validate the model (When this step is over, the model has finished the building and training step). Cross validation is the most frequently used method for learning and validation of the datasets. (d) Application of the final model in new data.

There are several machine learning methods which widely used for the study of cancer prediction and prognosis. They include (i) Artificial Neural Networks (ANNs), (ii) Decision Trees (DTs), (iii) Support Vector Machine (SVM), (iv) naive Bayesian and ect.

ANNs can be applied to automatic recognition, description and classification [23, 24]. ANNs model can be trained to get the relevant output after we give the model certain amount of input variables. As Figure 1 shows, ANNs contain input layer, output layer and multiple hidden layers between the input and output layers. The hidden layers represent the neural connections. At present, there has not been very good algorithm to determine the number of hidden layers [25]. Usually, it is based on experiences which are typically used for the process. The ANNs model can be characterized as a ’black-box’ technology. Trying to figure out the ANN working process or why it does not work is still a challenge.

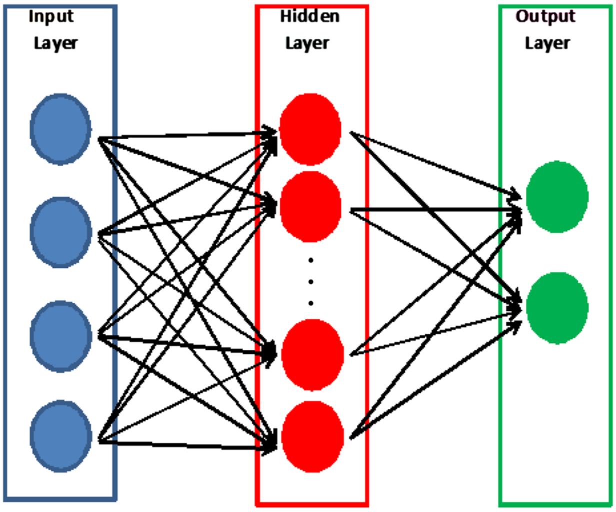

Figure 1: An illustration of the ANN structure. The arrows connect the input nodes to the output nodes. ANN: Artificial Neural Networks.

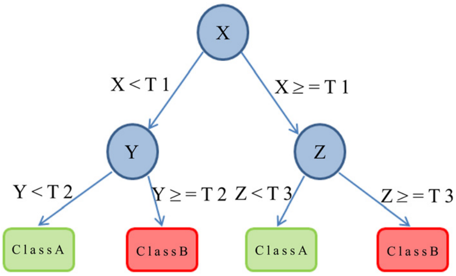

Figure 2: An illustration of a DT structure. The variables X, Y, Z are represented by a circle and the Class A and B are represented by squares. T1, 2, 3 represent the classification rules to classify the variables.

DTs follow a tree structure where nodes represent the input data and the leaves represent the output [26]. Figure 2 demonstrates a decision tree structure. The nodes represent the variables and the arrows represent the decision rules. Based on the tree structure, the traversing and classification are very quick. When there is a new sample, traversing the tree can allow sufficient reasoning. The specific architecture can help for conjecture and make classification decision. The approach is a prominent machine learning method and widely used in classifications [27].

Naive Bayesians are widely applied in classification, knowledge representation and reasoning [28]. The approach used the probability estimations rather than predictions. Naive Bayesians are composed of mainly directed acyclic graph and the classifiers are based on the probabilistic approach [29]. For known conditional densities, the Bayesians decision rules assign the classifications with the maximum posterior probability to obtain the optimal classifiers. Figure 3 shows an illustration of a BN.

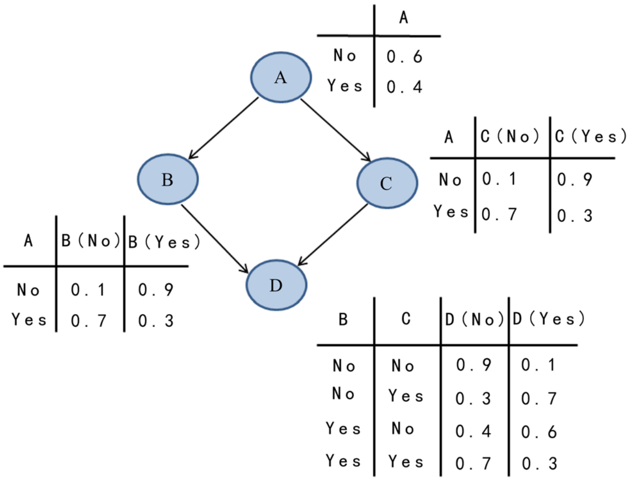

Figure 3: An illustration of a Naive Bayes (BN). Nodes A, B, C and D represent variables with their conditional probabilities.

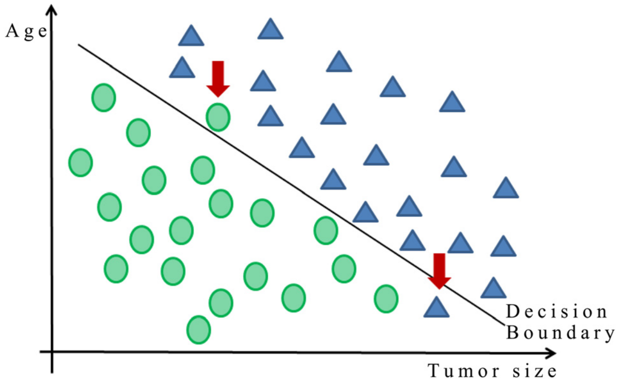

Figure 4: An illustration of a linear SVM classification. The classifications are according to the tumor size and patient age. The arrows are specified the misclassified tumors.

The nodes A, B, C and D are the variables and the calculated conditional probability (the edge) for each variable have been listed besides.

SVMs are the approach more widely used in cancer prediction and prognosis field recently [30, 31]. SVMs can map the input vector into high dimensional feature spaces. Then the hyperplane can be identified to separate the input data into two classes. The distance between the two hyperplanes (classes) should be maximized which can be used for reliable classifications. Figure 4 is a simple example which illustrates how a SVM model work to classify the tumor samples among benign and malignant tissues according to the conditions of age and tumor size.

These machine learning methods have often been used in cancer studies [32–35]. The algorithms are widely used to predict cancer types or classifications.

Machine learning methods in cancer prognosis and prediction

The ML techniques have been wildly used in cancer prognosis and prediction [36–39]. Most of the studies employed ML techniques for modeling cancer progression and identifying informative elements on classifications [40]. Gene expression profiles and clinical data have been used as the input data on the prognostic procedure. ML techniques have been used to predict cancer susceptibility, recurrence and survival.

The study in Ref. [25] used ANNs to discriminate between benign type and malignant type of breast cancer accurately and predict the probability of breast cancer for individual patients. In their work, ANNs are employed to be a prediction model which can classify malignant mammographic results from benign. Their dataset consisted of 62,219 consecutively collected mammography results. The dataset was fed as the input of the ANN model. The ANNs model is a 3-layer feed-forward ANN with 1,000 hidden layer nodes. There is a large number of hidden layer nodes in the models, as they have tested many times to have the results that large number hidden layer nodes generalizes better than networks with small number of hidden layer nodes. The authors trained and tested the model with 10-fold cross validation. They calculated AUC (Area Under Curve) to do the sensitivity analysis, the result is 0.965. At last, the authors made conclusion that their model can accurately discriminate malignant samples from benign ones and effectively predict the risk of breast cancer for individual abnormalities.

The study in Ref. [41] used an SVM model to predict the recurrence within 5 years of breast cancer. The data was a 679 patients dataset. Many types of variables are considered in the model such as histological grade, tumor size, number of metastatic lymph node, estrogen receptor, lymphovascular invasion, local invasion of tumor, and number of tumors and so on. The authors used the ML model to classified these patient data into high and low risk groups. Three prediction models: SVM, ANNs and Cox-proportional hazard regression model, were constructed and compared with each other in the study. The results show that the most effective model was SVM after comparing with other two established prognostic models. Another study in Ref. [37] used a decision tree system to study the recurrence of Oral squamous cell carcinoma (OSCC). The authors integrated a multitude of data such as clinical, imaging, and genomic ones. They identified the factors which dictate OSCC progression and predict potential recurrence of OSCC.

The study [42] used a decision tree (DT) model to identify patients with significant prostate cancer on prostate biopsy. The features of the DT including the information of the age, prostate-specific antigen (PSA), digital rectal examination (DRE), volume of the prostate, and PSA density (PSAD). The classification and regression tree (CART) analysis was carried out. The model resulted in an 83.3% area under the receiver operating characteristic (ROC) curve and detected out 92 patients which have significant prostate cancer of 221 prostate patients. The model of DT can help to reduce unnecessary biopsies without missing significant prostate cancer.

The study [36] is a prediction model developed for evaluation of survival women that have been diagnosed with breast cancer. There are three machine learning models was carried out for the prognosis of breast cancer survivability: SVM, ANNs, and semi-supervised learning models. The data contains surveillance, epidemiology, and end results database for breast cancer. The authors made a conclusion that semi-supervised learning model performed well that physicians can easily employ without consuming much time and effort for parameter selecting of the model. The ease of use and rapid outcome of the prediction model may ultimately lead to an accurate and less invasive prognosis for breast cancer patients.

According to the recent studies, most of these publications make use more than one ML algorithms and integrate different data types from various data sources for the prediction/prognosis of cancer types [43]. The study [44] detected the influence of genetic polymorphisms on breast cancer which measured 98 single nucleotide polymorphisms (SNPs) distributed over 45 breast cancer relevant genes in 174 patients, comparing machine learning methods of SVMs, DTs, and naive Bayes. In most cancer studies, the data should have both normal and cancer samples for the difference analysis. The research data set contains 174 samples from patients who were all females and newly diagnosed invasive breast cancer and 158 control samples who were anonymous females. The researchers used the whole data set which contains the patients with some missing SNP calls, the naive Bayesian method's accuracy can reach to 63% (the baseline is 50%). When the researchers prune the data set down to only complete patient genotypes, the naive Bayesian method's predictive power was increased to 67%. Generally, data preprocessing can effectively improve the prediction accuracy of the model. Besides keeping data integrity, noise reduction and de-noising are also often used methods. When the researchers provide a quadratic kernel SVM methods, the predicted accuracy further improves to 69%. The DTs with maximal performance achieves 68% ± 1% accuracy. The comparison can be seen in Table 1. Compared with other two models, the DTs have more balanced errors (The difference between sensitivity and specificity was smaller than that from other models), and the errors occur more evenly when predicting cancer and noncancerous patients. The prediction powers of different models also depend on the types of data and the mathematical problems to be predicted. In this data set and problem, the three learning algorithms naive Bayesian, SVMs and DTs performed similarly.

Table 1: Discrimination of breast cancer patients from normal controls using machine learning techniques. The mean and SD of five 20-fold cross validation trials.

Algorithm	Maximal accuracy (%)	Sensitivity (%)	Specificity (%)
Navie Bayes	67 ± 2	54 ± 2	79 ± 2
DTs	68 ± 1	67 ± 2	70 ± 4
SVM linear kernel	62 ± 2	57 ± 2	57 ± 2
SVM quadratic kernel	69 ± 4	53 ± 2	83 ± 7
SVM cubic kernel	67 ± 4	47 ± 2	84 ± 4

SVM, Support Vector Machine; DTs, Decision Trees.

Gene regulatory network identification methods in cancer biology

Cancer is a system disease and often related to polygene mutation and regulation variations. Therefore, understanding how these genes activate or repress of other genes has long been a goal of cancer biology. Detecting gene interactions or regulations by experiments can be very difficult and time consuming. There are approximately 20,000 genes in human genome. It is not feasible to check the gene interactions or regulations pair by pair. Researchers use DNA microarray [45], next-generation sequencing (most notably RNA-Seq) [46, 47] to a quantitative detect the transcriptomic profile of an individual cell or cell population. There are several often used methods to identified corresponding gene regulatory network such as Correlation, Regression and Bayesian methods (simple and dynamic Bayesian) etc.

Correlation is the most simple and basic method of network inference. Researchers often calculated the correlation for each pair of genes to know which genes are related. Kendall rank correlation coefficient and spearman correlation coefficient are the most often used methods. If the correlation coefficient value between the two genes' expressions reaches a certain value, then there is a certain correlation between these two genes and they can be linked together. This is useful for us if one wants to have a general idea of which gene pair is related. Since the connection has no direction or distribution between direct or indirect regulation, it is fast and scalable for large dataset. This kind of network is often named as a "gene co-expression network" rather than a gene regulatory network. Although the network identification is simple, the correlated networks can generate powerful results when appropriate analytical tools are applied. Weighted Gene Co-expression Network Analysis (WGCNA) is a widely used method in gene network studies. This tool can be used to calculate the correlations as the weights of the links for the network [48]. Recently, studies based on gene co-expression network [49–52] often combine protein-protein interaction (PPI), gene ontology (GO), quantitative real time polymerase chain reactions (qRT-PCR) or pathway-enrichment analysis to identify the significance of the gene modules or the biological functions of the identified dysregulated genes.

Regression is another common method in network identifications. This analysis method is more computationally expensive than the Correlation methods. It can provide the advantage of predicting causal direction. The regression method's simplest form follow the linear regression equation as below:

$$X_j = \beta_0 + \beta_1 X_i + \epsilon. \tag{1}$$

In Eq. (1), β_0 represents the intercept, β_1 represents the slope, the random error item is ε. β_1 can describe the relationship between X_j and X_i. It can be assigned as the weight of the edge from X_i to X_j. This is the simplest regression form. Its limitation is that linear relationships between genes are usually assumed. This will result to failure to detect non-linear regulations. A method named the Least Angle Regression (LARS) [53] made a progress to detect both linear and pre-defined non-linear regulations.

Bayesian methods are divided into simple and dynamic Bayesian ones. These methods have been used in gene network identification. When one use Bayesian method, a conditional probability $P(X_j|X_i)$ is often used to represent an interaction from gene *j* to *i*. X_j and X_i represent the gene expression level of gene *j* and *i*. Gene *j* can be seen as the parent of gene *i*. The graphical representation of these conditional probability can lead to the Bayesian networks. When one has gene expression data such as DNA microarray, one can use a maximum likelihood estimation to determine the highest posterior probability of the observed data. The Bayesian methods can give the direction of the edge of the network. Furthermore, Bayesian network can be regarded as a network of interactive prior knowledge. Therefore, the computational cost of Bayesian is higher than the methods have mentioned before.

Werhli and Husmeier [54] integrated gene expression data with multiple sources of prior biological knowledge to reconstruct the gene regulatory network. Moreover, Markov Chain Monte Carlo (MCMC) was used as a scheme to sample the hyperparameters from the posterior distribution. This can reduce reconstruction error as the values of the hyperparameters were close to be optimal. They have evaluated the results on the Raf related pathway. Including prior knowledge is advisable to improve the accuracy of the gene regulatory network reconstruction results.

Tan and Mohamad [55] proposed Bayesian network to reconstruct gene regulatory networks. Hill-climbing algorithm and Efron's bootstrap approach were applied with the Bayesian network model. At first, they use *Saccharomyces cerevisiae* cell-cycle gene expression dataset and *Escherichia coli* dataset because one can handle the microarray datasets with missing values. Then Bayesian networks using hill-climbing algorithm and Efron's bootstrap approach were applied. At the end, the result of the gene networks using *S. cerevisiae* dataset not only have achieved more than 90% positive prediction rate for the existing interactions and regulations, but the networks also have discovered potential interactions between genes.

Dynamic Bayesian can detect a feed-back loop and self-activated edge which is common in gene regulatory network. And that requires to be provided a gene expression variations at different specific time point. In practice, available data through DNA microarray, PCR and RNA-seq are not real time ones. The cells are dead for analysis each time point. Therefore, each cell can provide only one time point information. Although the real time high through put gene expression data do not exist, one may explore the ensemble of cell at each time point and treated as pseudo time series for studying the ensemble evolution. A temporal dataset in real time can allow one to detect a feed-back loop and self-activated edges and predict observations at a future time-step data. So obtaining temporal data or pseudo-temporal data are necessary for reconstructing a Dynamic Bayesian network (DBN). There are certain difficulties and complexities for the experiments to obtain this kind of data. DBN construction are more computational expensive than the simple Bayesian. Although DBN has many requirements of data set, it is still widely used as the network contains direction and feed-back loops. The Comparison of these methods can be seen in Table 2.

Dojer et al. [56–60] extended the framework of dynamic Bayesian networks to incorporate perturbations. Moreover, they use time series data from perturbation experiments and a discretization method was applied to infer an optimal network. The expressions of genes often have many much perturbations from some special treatments such as knockout experiments, this results the changes of the network interactions. The perturbations can be incorporated in the differential equation of mRNA and genes for the model. Based on the results, they show that the quality of inferred networks can markedly improve the accuracy due to the perturbed expression data.

Vinh et al. [61] demonstrated that using Dynamic Bayesian network (DBN) to reconstruct the biological network has limitations which only fit small sized networks. Furthermore, the DBN learning with local search or stochastic global optimization only can locate sub-optimal solutions. To overcome above defects, they integrated the DBN approach with a deterministic global optimization for genetic network construction using time course gene expression data. The proposed approach employed mutual

Table 2: Comparison of network construction methods.

Algorithm	Temporal data required?	Direction?	Applicable conditions	Instances
Correlation	No	Undirected	Large scale dataset; fast	[56]
Regression	No	Directed	No feed-back loops	[57]
Simple Bayesian	No	Directed	Fit small network; No feed-back loops	[58]
Dynamic Bayesian	Yes	Directed	Fit small network; detection of feed-back loops	[59]

information test (MIT) which is a novel scoring metric based on information theory for learning a global optimization structure. The GlobalMIT can learn high-order time delayed genetic interactions. A scoring function was applied to assess the goodness-of-fit of the DBN which can help to obtain optimal gene regulatory networks. As a result, they concluded that deterministic global optimization approaches can infer large scale genetic networks.

In Table 2, we have compared the advantages and disadvantages of the above models and some examples. We can see that the correlation method can process large scale dataset but the edges of the network have no directionality. The regression and simple Bayesian methods can construct the network with directions but do not have feed-back loops. The network constructed by dynamic Bayesian methods do have directions and feed-back loops but required temporal data.

The study [62] established a theoretical framework to analyze how the gene activity modulates metabolic pathway activity. They constructed the metabolic regulatory network through literature survey. They used a parameter randomization approach to identify the robust stable metabolic states of the regulatory network. They collected the key genes and related genes which are in the downstream of the key genes. Then they utilized unsupervised hierarchical clustering analysis (HCA) to determine the patterns of three clusters (W[Warburg], O[OXPHOS] and hybrid W/O). Through principal component analysis (PCA), they visualized the three clusters by projecting onto the first and second principal components. These results show that cells in the W state mostly use glycolysis for ATP production, cells in the O state mainly use OXPHOS for ATP production, and cells in the hybrid W/O state can utilize both glycolysis and oxidative phosphorylation (OXPHOS) to generate ATP. They pointed out targeting both OXPHOS and glycolysis may be necessary to eliminate cancer aggressiveness. This study offers a new perspective on coupling of gene regulation with metabolic pathways.

As biological events, such as transcription, translation, biochemical reactions, occur at multiple time scales, approaches with only data mining and machine learning may result in ill-posed problems or non-physical solutions. The study [63] discussed how machine learning and multi-scale model can be integrated to reach a ‘synergy’ effect. This paper reviewed how machine learning methods which provide the appropriate tools can train data, prevent over-fitting, and identify correlations etc. In the mean time, the multi-scale models can integrate the underlying physics or biological features to explore the interactions, related mechanisms and understand the function of the predicting system dynamics and causality. This study can provide new insights into disease mechanisms, help us to identify new targets on treatment strategies and make decisions which are beneficial for human health.

Using gene regulatory network to study cancer

Non-linear dynamical models from ordinary differential equations

ODE models are the effective analysis approach for non-linear dynamical systems as they can be used to describe the gene regulatory network dynamics, multi-stability, limit cycles, chaos and etc. ODE models are more quantitative because they represent the underlying physical system dynamic by continuous variables while others often use discrete variables. Furthermore, the extensive experimental literature can provide options on ODEs models, such as parameter selections and kinetic laws in functions [64].

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_i\left(x_1, x_2, \ldots x_n, p, u\right). \tag{2}$$

In Eq. (2), x is the expression level of gene i at time t. n represents the gene number and u represents an external perturbation of the system. F_i is a vector which represents the deterministic driving force. In ODE models, continuous time variables are used and there is no negative value as the protein and mRNA productions cannot be negative. The degradation rate of mRNA or proteins is assumed to be a constant to simplify the computations.

In the study [65], the researchers constructed a regulatory network which consists of regulatory proteins and metabolites. After an extensive literature analysis, a comprehensive network was constructed capturing the regulations of oxidative respiration and glycolysis on both genes and metabolites. The network contains two major metabolic pathways: aerobic glycolysis, glucose oxidation, fatty acid oxidation and glutamine oxidation (OXPHOS). These two pathways inhibit each other as they compete for shared metabolites. The network also contains the specific genes which directly regulate the pathways. For example, the fatty acid and glucose oxidation are regulated by AMPK which is an energy sensor gene and the glycolytic pathway is regulated by HIF-1 which is a hypoxia inducible factor. Metabolites such as ROS produced by the pathways are also reflected in the network. Some regulated genes such as RAS, MYC, and c-SRC coupled to oncogenic pathways are also included in the network.

The researchers have coarse-grained gene regulatory circuit into a minimal circuit which are AMPK, HIF-1, and ROS shown in Figure 5.

This can help us to understand the major behaviors of the network. The effects of other genes can be regarded as the input of the network. The dynamical behaviors of the gene regulatory network (AMPK, HIF-1, ROS) can be quantified by the non-linear differential rate equations as below [65]:

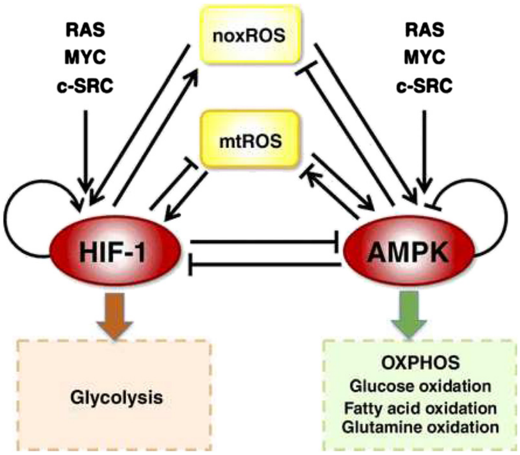

Figure 5: The coarse-grained the network of AMPK:HIF-1:ROS regulatory circuit. ROS represents both mtROS and noxROS. RAS, MYC, and c-SRC modulates the balance of glycolysis and OXPHOS (Image source: Yu et al. [65] with permission). noxROS, NADPH oxidase reactive oxygen species; mtROS, mitochondrial reactive oxygen species; HIF-1, hypoxia-inducible factor 1; AMPK, AMP-activated protein kinase; OXPHOS, oxidative phosphorylation.

$$\frac{\mathrm{d}X}{\mathrm{d}t} = g_X \cdot G - X \cdot k_X \cdot K. \tag{3}$$

In Eq. (3), X represents the gene expression level of the protein. g_X and k_X are the production and degradation rates of gene X, respectively. G and K represent the regulation functions of the production and degradation of X, respectively. The authors used a non-linear function named Hill function to describe the regulation of X by a component Y.

$$H^{s}\left(Y, Y^{0}, \lambda_Y, n_Y\right) = H^{-}\left(Y, Y^{0}, n_Y\right) + \lambda_Y H^{+}\left(\left(Y, Y^{0}, n_Y\right).\right. \tag{4}$$

Here, Y^0 is the threshold level of Y, n_Y represent Hill coefficient and λ_Y is the maximum or minimum fold change of X due to the regulation of Y. In Eq. (4), H^- and H^+ are the inhibition Hill function and excitatory Hill function, respectively which are shown as Eqs. (5) and (6).

$$H^{-}\left(Y, Y^{0}, n_Y\right) = \frac{1}{\left[1 + \left(\frac{Y}{Y^{0}}\right)\right]^{n_Y}}. \tag{5}$$

$$H^{+}\left(Y, Y^{0}, n_Y\right) = \frac{\left(\frac{Y}{Y^{0}}\right)^{n_Y}}{\left[1 + \left(\frac{Y}{Y^{0}}\right)\right]^{n_Y}}. \tag{6}$$

λ_Y<1 represents an inhibitory regulation and λ_Y>1 represents an excitatory regulation. The regulatory function of production X is $G = H^s(Y, Y^0, \lambda_Y, n_Y)$; The regulatory function of degradation X is $K = H^s(Y, Y^0, \lambda_Y, n_Y)$, here X is regulated by Y. When X is regulated by two components Y and Z simultaneously, the function of production or degradation can be written as Eq. (7):

$$G\,(\mathrm{or}\,K) = \begin{cases} H^{s}\left(Y, Y^{0}, \lambda_Y, n_Y\right) H^{s}\left(Z, Z^{0}, \lambda_Z, n_Z\right) & \text{Y and Z are independent} \\ C^{\mathrm{comp}}\left(k_0, Y, Y_X^{0}, k_Y, n_Y, Z, Z_X^{0}, k_Z, n_Z\right) & \text{Y and Z are competitive} \end{cases}. \tag{7}$$

Based on the ODEs, the two stable steady states emerged [65]. In normal cells, there are two steady states: one is from the Warburg effect (W) and the other is from the oxidative respiration (O), shown in Figure 6A. This result is consistent with the fact that cells usually use glucose oxidation to produce energy, but during the anaerobic exercises, cells turn to glycolysis. Then the analysis for the cancer cells are reflected by larger γ and lower k_h. It was found that cancer cells have a new hybrid state (W/O) shown in Figure 6B [65]. In the hybrid state (W/O), the expression levels of pAMPK and HIF-1 are both high. this can enhance metabolic plasticity to promote tumor occurrence and metastasis. The hybrid state illustrates that the cancer cells have the capability to

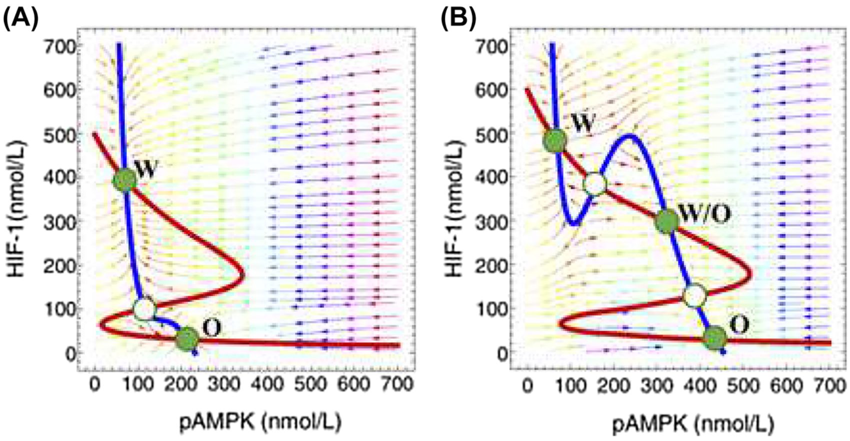

Figure 6: The steady states in the phase space of AMPK and HIF-1. The red and blue lines represent the null cline of embedded image and the null cline of embedded image, respectively. The green solid and green hallow dots denote stable steady states and unstable steady state, respectively (Image source: Yu et al. [65] with permission). PAMPK, phosphorylated AMP activated protein kinase; HIF-1, Hypoxia-inducible factor 1.

simultaneously make use of both glycolysis and mitochondrial OXPHOS.

The researchers found eight types of cancer from TCGA database to evaluated the results and further proposed to design metabolic therapies by considering the hybrid metabolic phenotype (W/O) [65]. By simulating the effectiveness of several therapies through the model, they found that metformin is more likely to push the cells into the hybrid metabolic phenotype than the others. This model can also give explanations of why hybrid therapies are more effective [65].

Researchers [66] pointed out that if one wants to use the immune system as a treatment for cancer, the contribution of tumor microenvironment to the complex interactions between the cancer cells and the immune response must be better understood. Exosomes are small (30 – 200 nm) vesicles whose functions are to transfer proteins, mRNAs, and microRNAs to other cells. Researchers developed an ODEs model to study the role of exosome communication which are critical to the cancer immunity interplay. This model can help researchers compare the effectiveness of radiation therapy to the combination with immune therapy and illustrate how the model can provide basis for the design and evaluation of the combination therapies.

As the parameter determination of the model will consume much time, researchers developed a computational method named RACIPE which can reduce the searching time during model parameter determination of ODE models [67]. The input of RACIPE is the topology of gene regulatory network, and generates an ensemble of models with random kinetic parameters. Then conventional ODE based simulation is used for each random model to obtain steady states. Finally, relevant analysis is carried out according to the corresponding biological characteristics on the gene expression data from all models. This can help to pin down the range of the parameters consistent to the global behaviors.

The landscape flux models through stochastic differential equations

The landscape flux theory in non-equilibrium system

In cellular environment, the transcriptional, translational, and post-translational regulations due to either low copy number of molecules or slow switching among the states of promoter structure, chromatin epigenetics, or nuclear architecture can lead to intrinsic noise or fluctuations [68]. The pathway-specific or global differences in the abundance of cellular components, or differences in the timing of cell-cycle events, environments can lead to extrinsic noise or fluctuations [69, 70]. The intrinsic and extrinsic noise cannot usually be ignore. Compared to the ODE model, a stochastic driving force which can describe the effects of both the intrinsic and extrinsic noise can be introduced into the stochastic dynamic model (stochastic differential equation (SDE) model). The stochastic dynamics of the system can then be described by the Langevin equation [71–73]:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = F_{\mathrm{i}}(x) + \eta(x,t). \tag{8}$$

For the stochastic dynamics described by the Langevin equation, researchers can obtain the probability distributions by collecting the real time trajectories. In Eq. (8), $\eta(\mathbf{x}, t)$ is the stochastic force obeying Gaussian distribution and $\langle \eta(\mathbf{x}, t) \rangle = 0$ and $\langle \eta(\mathbf{x}, t)\eta^T(\mathbf{x}, t') \rangle = 2\mathbf{D}(\mathbf{x})\delta(t - t')$. $\mathbf{D}(\mathbf{x})$ is a diffusion matrix which characterizes the fluctuation strength and the correlation.

Since the individual stochastic trajectories are unpredictable, the probability evolution is linear and predictable. The corresponding probability ($P(\mathbf{x}, t)$) evolution equation follows the Fokker–Planck equation: $\partial P(\mathbf{x}, t)/\partial t = -\nabla \cdot [\mathbf{F}P - \nabla \cdot (\mathbf{D}P)]$. This can be written as $\partial P/\partial t = -\nabla \cdot \mathbf{J}$. The meaning of this equation is a local probability conservation law. The local probability change is equal to the net in or out flux. The flux is given as $\mathbf{J} = \mathbf{F}P - \nabla \cdot (\mathbf{D}P)$, where the first term is related to the deterministic force while the second term represents the fluctuation. In the steady state, $\partial P_{ss}/\partial t = \nabla \cdot \mathbf{J}_{ss} = 0$, where $\mathbf{J}_{ss}$ stands for steady state. The probability flux of steady state satisfies $\mathbf{J}_{ss} = \mathbf{F}P_{ss} - \mathbf{D} \cdot \nabla P_{ss}$. When $\mathbf{J}_{ss}$ is deviated from zero, the detailed balance is broken and the degree of non-equilibrium away from equilibrium can be quantified. Since the steady state flux $\nabla \cdot \mathbf{J}_{ss} = 0$. This indicates that the steady state flux is rotational. Therefore, for non-equilibrium systems, the driving force $\mathbf{F}$ can be decomposed into a gradient of the potential landscape and a curl flux force [74, 75]: $\mathbf{F} = -\mathbf{D} \cdot \nabla U + \mathbf{J}_{ss}/P_{ss}$, where $U = -\ln P_{ss}$ can be viewed as the potential landscape of the system [76]. In general, the gene regulatory networks often have more than two genes. It is impossible to view the landscape in more than two or three dimensions. Researchers usually projected the landscape into two or three chosen genes or collective coordinates from the combination of different genes such as principal component analysis (PCA) to view the landscape clearly.

Effects on biological noise

In biological system, the noises mainly involve intrinsic and extrinsic noise. The intrinsic noise often refers to statistical number fluctuations in a given process such as DNA transcription into RNA, RNA translation into a peptide, a peptide folding into a functional protein and protein degration etc. The extrinsic noise usually refer to the fluctuations from environments. In the SDEs, the term η is used to represent the noise amplitude. If the system only contains few molecules, intrinsic noise will be the primary consideration. If there are large number molecules in the system, intrinsic noise is usually small and extrinsic noise can be significant, as the extrinsic noise is due to environmental factors associated with other processes. This can be described by the SDEs as the set of equations which can describe the system dynamics.

One can quantify the degrees of noise or fluctuations by calculating Fano factor. Fano factor is defined as: $F = \sigma^2/\mu$ which represents noise strength. Here, σ is the standard deviation and μ is the mean of the probability distribution. Ideally, the independent stochastic processes is expected to be a Poisson distribution. In that case, the value of Fano factor is 1. If the value of Fano factor is large (small) compared to 1, it implies the degree of noise or fluctuations is large (small) and the underlying statistical distribution is deviated from Poisson.

Self-consistent mean field approximation

Usually, solving the Fokker–Planck equation to get the time dependent and the steady state probability/potential landscape is difficulty. The self-consistent mean field approach [72] can provide an approximation by assuming a separable form of the probability distribution $P(x_1, x_2, \cdots, x_n, t) \sim \prod_i P(x_i, t)$. Therefore, the probability can be solved self-consistently. The dimensionality in the problem is reduced from m^n to $m \times n$, making the computation more tractable.

The Gaussian Probability Distribution can be used as an approximation for the form of the probability distribution. For small fluctuations, the mean vector $\bar{\mathbf{x}}$(t) and the covariance matrix $\sigma(t)$ of the Gaussian distribution obey the following moment equations:

$$\dot{\bar{\mathbf{x}}}(t) = \mathbf{F}(\bar{\mathbf{x}}(t))\dot{\sigma}(t) = \mathbf{A}(t)\sigma(t) + \sigma(t)\mathbf{A}^T(t) + 2\mathbf{D}(\bar{\mathbf{x}}(t)). \quad (9)$$

The elements of the matrix $\mathbf{A}$ are given by $\mathbf{A}_{ij}(t) = \frac{\partial F_i(\bar{\mathbf{x}}(t))}{\partial \bar{x}_j(t)}$. Due to the self-consistent mean field approximation of separable distributions, only diagonal elements of $\sigma(t)$ are considered in Eq. (9). Thus based on the approximation of separable Gaussian distributions, the evolution of the probability distribution to each variable x_i is given by

$$P(x_i, t) = \frac{1}{\sqrt{2\pi\sigma_i(t)}} \exp\left\{ -\frac{\left[x_i - \bar{x}_i(t)\right]^2}{2\sigma_i(t)} \right\}. \quad (10)$$

For a monostable system, the steady state probability distribution obtained from Eq. (10) is a separable Gaussian distribution centered at the fixed point. For a multistable system, there is a separable Gaussian distribution associated with each fixed point. The final steady state probability distribution $P_{ss}(\mathbf{x})$ is constructed as a linear combination of these Gaussian distributions, with the combination coefficients chosen to be the relative frequencies of occurrence of the corresponding fixed points obtained by running different initial conditions.

Optimal path through path integral formulation

Consider stochastic systems governed by the Fokker–Planck equation with a constant diffusion matrix: $\partial P(\mathbf{x}, t)/\partial t = -\nabla \cdot [\mathbf{F}(\mathbf{x})P(\mathbf{x}, t) - \mathbf{D} \cdot \nabla P(\mathbf{x}, t)]$. Based on the Onsager-Machlup functional approach [76], the transition probability from the initial state $\mathbf{x}_{\text{ini}}$ at time t_i to the final state $\mathbf{x}_{\text{fin}}$ at time t_f is given by a path integral: $P(\mathbf{x}_{\text{fin}}, t_f\,; \mathbf{x}_{\text{ini}}, t_i) = \int \mathscr{D}[\mathbf{x}(t)] \exp\{-S[\mathbf{x}(t)]\} = \int \mathscr{D}[\mathbf{x}(t)] \exp\{-\int \mathscr{L}(\mathbf{x}(t))\mathrm{d}t\}$, where $\mathscr{L}(\mathbf{x}(t)) = \frac{1}{4}(\dot{\mathbf{x}} - \mathbf{F}(\mathbf{x})) \cdot \mathbf{D}^{-1} \cdot (\dot{\mathbf{x}} - \mathbf{F}(\mathbf{x})) + \frac{1}{2}\nabla \cdot \mathbf{F}(\mathbf{x})$ is the Lagrangian and $S[\mathbf{x}(t)] = \int \mathscr{L}(\mathbf{x}(t))\mathrm{d}t$ is the action of the path. The notation $\int \mathscr{D}[\mathbf{x}(t)]$ represents an integral over all the possible paths beginning from the initial state $\mathbf{x}_{\text{ini}}$ at time t_i and ending in the final state $\mathbf{x}_{\text{fin}}$ at time t_f. According to this formula, each path is assigned with a probability weight, $\exp\{-S[\mathbf{x}(t)]\}$, associated with the action of that path. The dominant or optimal kinetic paths are identified as the paths with the maximum probability. In non-equilibrium systems the non-vanishing curl flux $\mathbf{J}_{\text{ss}}$ drives the kinetic path to deviate from the steepest descent path on the landscape. Therefore, the kinetic paths of the non-equilibrium systems are in general irreversible.

Non-equilibrium thermodynamics

From the Fokker–Planck equation, we can obtain the intrinsic entropy of the stochastic non-equilibrium dynamical system as $S = \int P(\mathbf{x}, t)\ln P(\mathbf{x}, t)\mathrm{d}x$. In the non-equilibrium system, exchange in energy and information results the dissipation. It depicts a global physical characterization of the non-equilibrium system. The change of the system entropy with time can be written as: $\dot{S} = \dot{S}_{\text{t}} - \dot{S}_{\text{e}}$, $\dot{S}_t = \int \mathrm{d}\mathbf{x}(\mathbf{J} \cdot (D\mathbf{D})^{-1} \cdot \mathbf{J})/P$, where $\dot{S}_{\text{t}}$ is the entropy production rate which is positive or zero and $\dot{S}_{\text{e}} = \int \mathrm{d}\mathbf{x}(\mathbf{J} \cdot (D\mathbf{D})^{-1} \cdot \mathbf{F}')$ is the heat dissipation or entropy flow rate which can be either positive or negative. $\mathbf{F}' = \mathbf{F} - D\nabla \cdot \mathbf{D}$ is the effective force. If we define $\dot{S}$ as the entropy change of the system of the non-equilibrium system, then $\dot{S}t$ is the total entropy change of the system and the associated environment. In the steady state, $\dot{S} = 0$, the heat dissipation is equal to the entropy production rate [77]. The flux of the system can be written as: $J(x, t) = \mathbf{F}P - D\nabla P$.

Applications of the landscape flux models

Recently, researchers used literature research and text mining methods to reconstruct gene regulatory network to study the cancer progression and metastasis [78–80]. Cancer is a fetal disease regulated by the underlying gene networks. The researchers use EVEX database to search the experimental literature and find which genes regulate which genes. The regulations of the network are all obtained from the experiments. The regulations not only have directions and feed-forward loops but also contain the genes promoted or suppressed by other genes.

Researchers in study [81] reconstructed a cancer gene network which contains 32 nodes (genes) and 111 edges (regulations) shown in Figure 7. The arrows represent activations and the filled circles represent the repressions. In

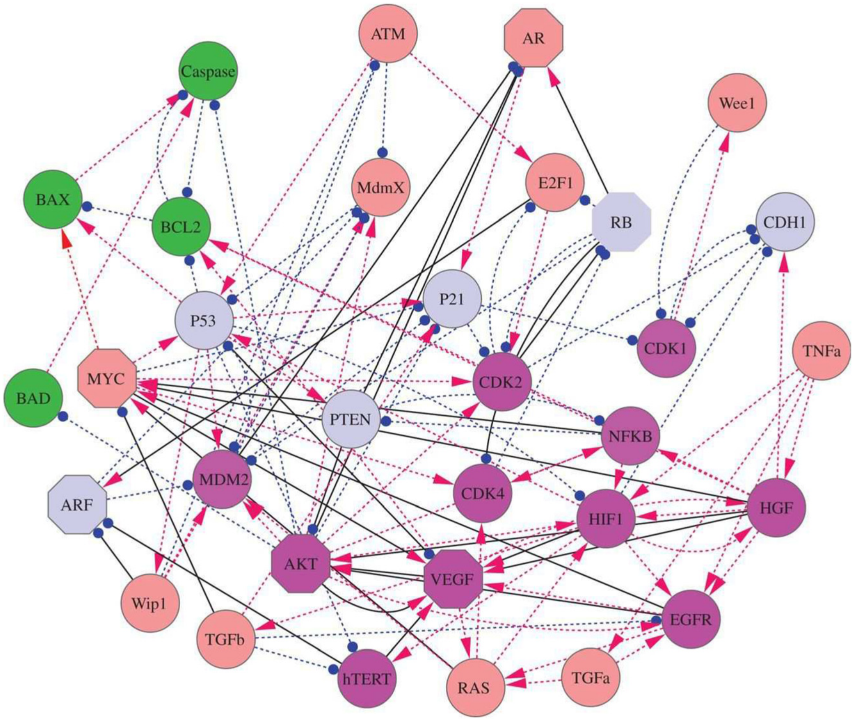

Figure 7: The regulation network of cancer including 32 nodes and 111 edges (66 activation regulations and 45 repression regulations) (Image source: Li et al. [81] with permission). VEGF, vascular endothelial growth factor; TGF, transforming growth factor; TNF, tumor necrosis factor; CKD, chronic kidney disease.

the gene regulatory network, the green nodes are apoptosis genes, the magenta nodes are cancer marker genes and the light blue nodes are tumor repressor genes. To describe the dynamics of the underlying network, the researchers developed the corresponding ordinary differential equations. For the Hill regulatory function term, the activation and repression can be represented. The equations are shown as below:

$$F_i = -K_i \times X_i + \frac{a \times X_{a_i}}{S^n + X_{a_i}} + \frac{b \times S^n}{S^n + X_{b_j}}. \tag{11}$$

In the Eq. (11), $i = 1, 2, \ldots, 32$, so there are 32 equations. S represents the threshold of the sigmoid function which is specified as $S = 0.5$. The Hill coefficient n determines the steepness of the sigmoidal function which is specified as $n = 4$. k is the self-degradation constant and a and b are the activation and the repression constants, respectively. To certain node i, X_{ai} and X_{bj} represent average interaction strengths for activation and repression from other genes. For each node i, X_{ai} is defined as $(X^n_{a(1)} \times M(a(1), i) + X^n_{a(2)} \times M(a(2), i) + \cdots + X^n_{a(m1)} \times M(a(m1), i))/m1$. X_{bi} is defined as $(X^n_{b(1)} \times M(b(1), i) + X^n_{b(2)} \times M(b(2), i) + \cdots + X^n_{b(m2)} \times M(b(m2), i))/m2$. Here, $a(1)$, $a(2)$, … $a(m1)$ is a list of node number which activate node i, and $b(1)$, $b(2)$, … $b(m2)$ is a list of node number which repress node i. $M(j, i)$ $(j, i = 1, 2, \ldots, 32)$ is a matrix acquired by multiplying the interaction type matrix M_i and interaction strength matrix M_s from node j to node i. $M(j, i) = M_\mathrm{i}(j, i) \times M_\mathrm{s}(j, i)$ $(i, j = 1, 2, \ldots, 32)$. In Eq. (11), the first term represents self-degradation, the second term represents activation and the third term represents repression.

The 32 ODEs can describe the driving force of the system and govern the network dynamics. According to $U = -\ln P_{\mathrm{ss}}$, the researchers quantified the potential landscape of the system. Figure 8 shows that the landscape contains three states characterized by the basins of attractions which are normal, cancer and apoptosis state, respectively. The landscape can reflect the cancer progression such as cancerization and apoptosis process. The landscape topography, the transition rate, and the dominant kinetic paths are determined both by the landscape gradient and curl probability flux as illustrate in Figure 8B. Together, they (Figure 8A and B) can give a global and system view of cancer progression and apoptosis process which can provide the physical explanation and quantification for the underlying mechanisms of cancerization.

Since the regulations can influence the landscape topography, the researchers use global sensitivity analysis to uncover the key regulations or genes in the gene network influencing the stability and topography of the cancer landscape. The results of global sensitivity analysis can provide a way to identify the key elements which determine the cancerization and apoptosis process. Moreover, the results can help the researchers making certain predictions about which regulations are crucial for cancer progression and cancer treatment.

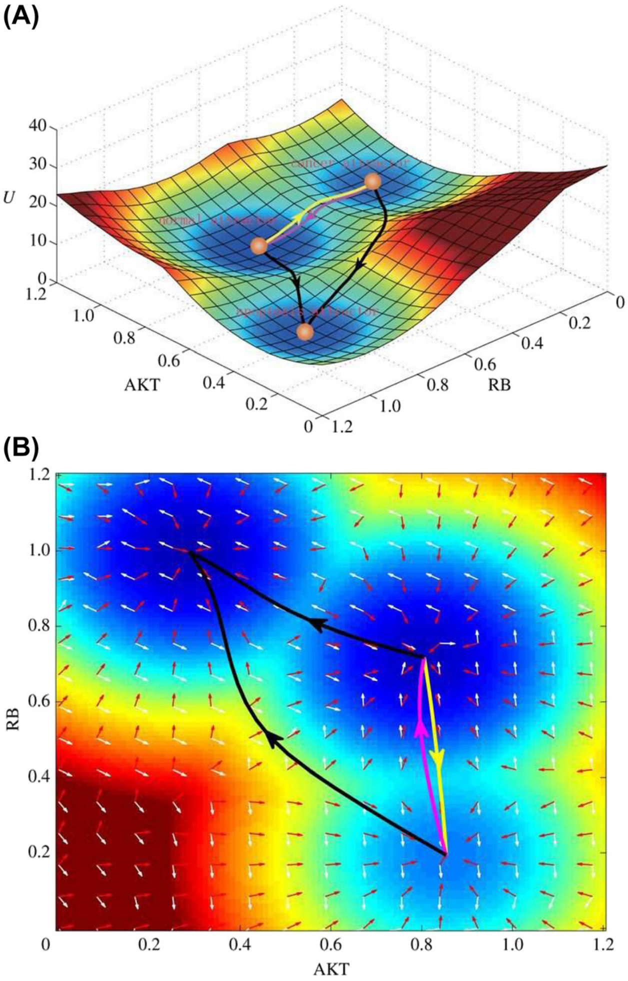

Figure 8: The tristable landscape for the cancer network.
(A) The three dimensional landscape and dominant kinetic paths. The yellow path, the magenta path and black paths represent the path from normal state to cancer state, from cancer state to normal state and from the normal and cancer state to apoptosis state, respectively. (B) The corresponding two dimensional landscape. Red arrows and white arrows represent the negative gradient of potential energy and the probabilistic flux, respectively (Image source: Li et al. [81] with permission).

The researcher used the stochastic dynamical model to further study specific types of cancer such as breast cancer and gastric cancer [79, 82]. Here, we summarized the gastric cancer as an example [82]. They identified a gene regulatory network to investigate gastric cancer. The gene regulatory network contains both the genetic level and epigenetic information of gastric cancer and gastritis. As

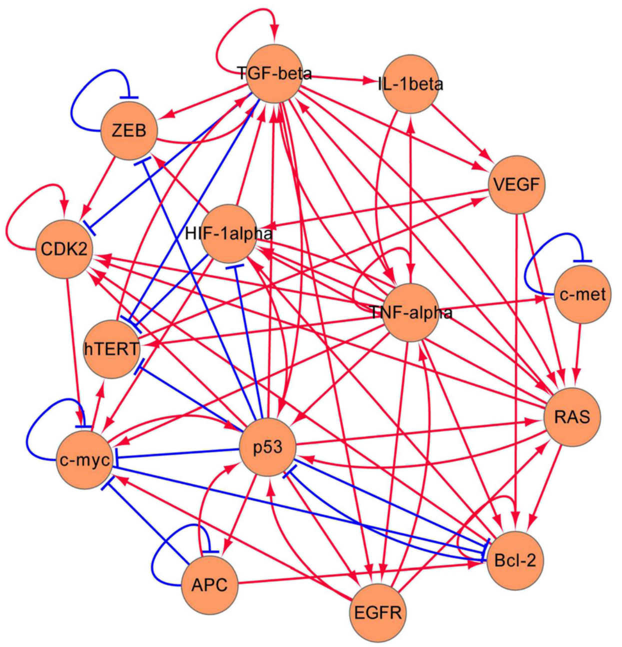

Figure 9: The regulatory network of the gastric cancer with 15 nodes and 72 regulations (57 activations and 15 repressions. The arrows represent the activating regulations and the short bars represent the repressing regulations) (Image source: Chong et al. [82] with permission). TNF-alpha, Tumor necrosis factor-alpha; HIF-1alpha, Hypoxia-inducible factor 1 alpha; TGF-beta, transforming growth factor-beta; ZEB, zinc-finger E-box-binding; CKD2, chronic kidney disease stage 2; APC, Adenomatous polyposis coli; EGFR, epidermal growth factor receptor; VEGF, vascular endothelial growth factor.

shown in Figure 9, there are 15 nodes (genes) and 72 edges (regulations).

The researchers then quantified the dynamics of the underlying gene regulatory network by the ODEs:

$$\frac{\mathrm{d}X_i}{\mathrm{d}t} = F_i = g_i \prod_{j=1}^{n_i} H_{ji} - k_i X_i \qquad (12)$$

In Eq. (12), $\frac{\mathrm{d}X_i}{\mathrm{d}t}$ represents the specific gene expression or protein concentration changes with time, g and k are the generation rate and self-degradation rate of the gene or protein, respectively. X_i represents the gene expression level (protein concentration) of the gene i. j denotes the gene which regulates the gene i. n_i is the gene number which regulates the gene i. H_{ji} is a Hill function [65] defined below:

$$H_{ji} = \frac{S_{ji}^n}{S_{ji}^n + X_j^n} + \lambda_{ji}^r \frac{X_j^n}{S_{ji}^n + X_j^n} \qquad (13)$$

In Eq. (13), S represents the "threshold" of the sigmoid regulatory function. n is the Hill coefficient for depicting the steepness of the sigmoid function which describes the cooperatively of the interactions. The parameter λ_{ji} denotes the regulation strength from X_j to X_i which is a real number greater than 1. r denotes the regulation type. If the regulation type is activation, r is set to be +1. If the regulation type is inhibition, r set to be −1. There are 15 genes, so the gene regulatory network can be described by 15 ODEs with additional noise term describing the intrinsic and external fluctuations. The researchers are able to quantified the potential landscape according to $U = -\ln P_{ss}$, where P_{ss} is the steady state probability function obtained by the collection of statistics of the stochastic trajectory simulation of the gene expression. In Figure 10, there are three stable states which are normal, gastritis and gastric cancer, respectively. The definition of each state is based on the

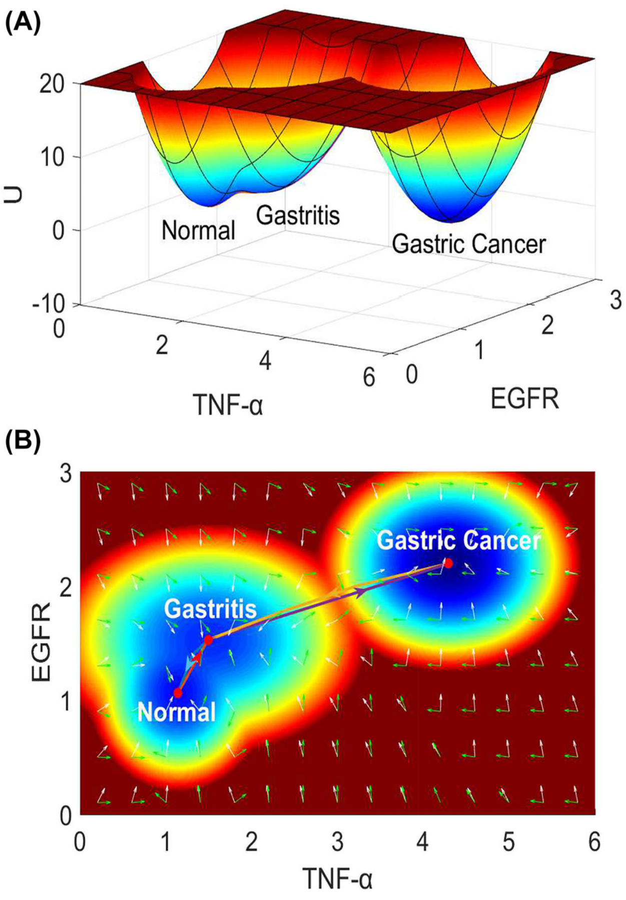

Figure 10: The tristable state landscape of the gastric cancer. (A) The three dimensional landscape and dominant kinetic paths. (B) The corresponding two dimensional landscape of the gastric cancer. The lines in red, blue, violet and yellow represent respectively the dominant kinetic path from the normal to the gastritis state, from the gastritis to the normal state, from the gastritis to the gastric cancer state, and from the gastric cancer to the gastritis state. White arrows and green arrows represent the negative gradient of the potential landscape and the steady state probability curl flux force, respectively (Image source: Chong et al. [82] with permission). TNF-α, Tumor necrosis factor-α; EGFR, epidermal growth factor receptor.

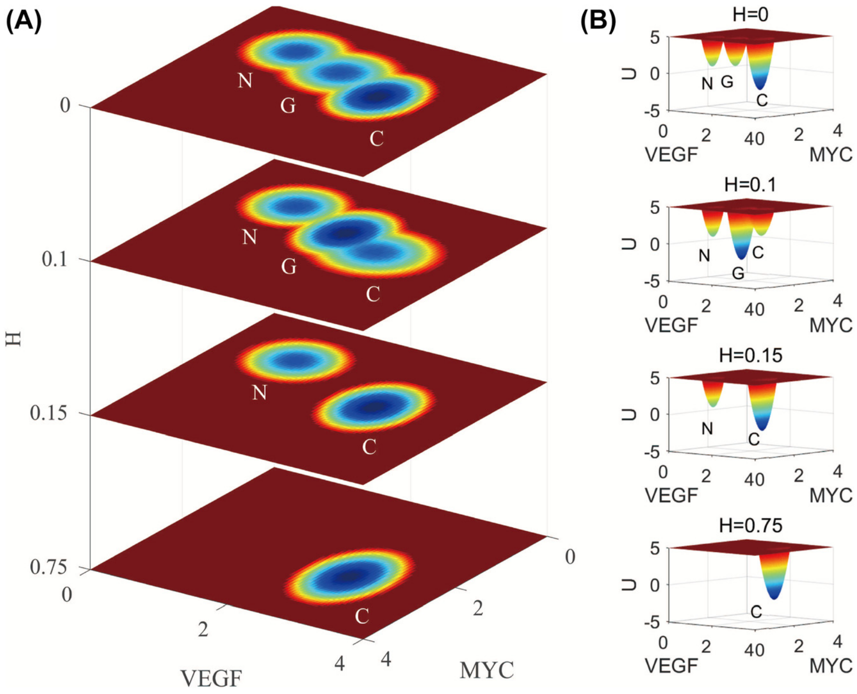

Figure 11: The changes of the landscape topography of the gastric cancer upon the influences of the *H. pylori* infection in two dimensions (A) and three dimensions (B). The two horizontal axes represent gene expressions while the vertical axis represents the degree of the *H. pylori* infection *H*. The meaning of *N*, *G* and *C* are normal, gastritis and gastric cancer, respectively (Image source: Chong et al. [82] with permission). VEGF, vascular endothelial growth factor.

biological function and gene expression levels. The lines in red, blue, violet and yellow are the dominant kinetic paths from one state to the other. The white arrows are the negative gradient of the potential landscape and the green arrows are the curl flux force. The barrier heights and dominant kinetic paths can reflect the global stabilities and the state switching routes which can quantified the gastric cancer formation process.

To investigate the underlying mechanism caused by genetic or epigenetic factors, the authors simulate how *Helicobacter pylori* (*H. pylori*) infection can influence gastric cancer progression [83, 84]. Figure 11 shows different *H. pylori* infection degrees can lead to the landscape topography changes which illustrates the process from the normal state through the gastritis to gastric cancer state. In Figure 11, the label *H* represents the *H. pylori* infection degrees. When the value of *H* is increased, the basin of the gastritis state becomes deeper and closer to gastric cancer state basin, and finally emerges to the gastric cancer state. This can demonstrate how the *H. pylori* accelerates the gastric cancer formation with gastritis.

Researchers in study [78] identified the gene-metabolism integrative network to quantify the global driving forces for cancer metabolism dynamics through the underlying landscape and probability flux. As shown in Figure 12, the gene-metabolism integrative network contains two parts: gene regulatory network and metabolic pathway. The gene regulatory network and the metabolic pathway are bridged by the gene-enzyme and metabolite-gene interactions. The genes (such as Akt, p53, cMyc, PTEN, HIF-1, and PDK), enzymes(such as G6P and F2,6BP), metabolites and interactions (such as lactate, ROS, ATP, and O_2) among them compose a cancer gene-metabolism integrative network.

The researchers quantified the driving force for the network dynamics of the gene expressions or enzyme concentrations by ODEs as below:

$$\dot{X}_i = F(X_i) = A_i \prod_{j=1}^{N_i} H_{ji} - D_i X_i \tag{14}$$

$$H_{ji} = \frac{S_{ji}^n}{S_{ji}^n + X_j^n} + \gamma_{ji} \frac{X_j^n}{S_{ji}^n + X_j^n} \tag{15}$$

$$= (\gamma_{ji} - 1) \frac{X_j^n}{S_{ji}^n + X_j^n} + 1 \tag{16}$$

$$= (1 - \gamma_{ji}) \frac{S_{ji}^n}{S_{ji}^n + X_j^n} + \gamma_{ji}$$

In Eq. (14), X represents the gene expression level or concentration of enzyme. F represents the driving force of the variable X. A and D represent the basic production rate and degradation rate of the gene or the enzyme, respectively. In Eq. (15), S is the Hill coefficient denoting the gene expression level with half threshold of production. n is the Hill coefficient which can represent the degree of the cooperativity of the interactions. $\boldsymbol{H_{ji}}$ is a non-linear function which is often named as shifted Hill function [66]. In

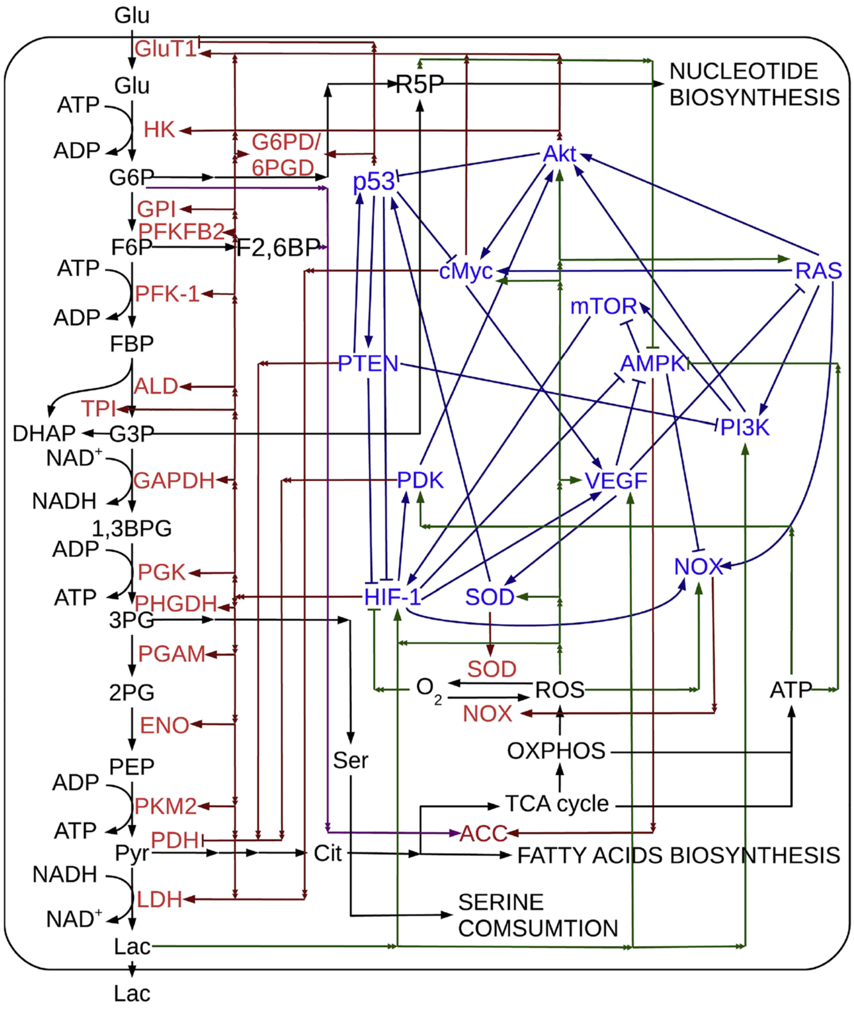

Figure 12: Cancer gene-metabolism integrative network. Genes are marked with blue. Enzymes are marked with red. Metabolites are marked with black. Dark blue arrows and bars represent gene-gene interactions. Dark red arrows and bars represent and gene-enzyme regulations. Purple arrows and bars represent metabolite-enzyme regulations. Black arrows represent biochemical reactions. Double arrows represent shared lines for multiple regulations. Each of the same colored connections that start with double arrows and end with solid arrows and bars represent one regulation (Image source: Li et al. [78]). Glu, glutamate; ATP, Adenosine triphosphate; FBP, filtered backprojection; ALD, atomic layer deposition; TPI, triosephosphate isomerase; DHAP, dihydroxyacetone phosphate; NAD,nicotinamide adenine dinucleotide; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; PGK, phosphoglycerate kinase; PHGDH, phosphoglycerate dehydrogenase; PGAM, phosphoglycerate mutase; ENO, enolase; PEP, positive expiratory pressure; PKM, Pyruvate kinase muscle; PDH, pyruvate dehydrogenase; OXPHOS, oxidative phosphorylation; PTEN, phosphatase and tensin homolog; VEGF, vascular endothelial growth factor.

Eq. (16), the parameter γ_{ji} represents the regulation type of X_i from X_j. If γ>1, γ_{ji} represents the activation and if γ<1, γ_{ji} represents the inhibition. The parameters for this cancer metabolism model are chosen carefully for producing the results that are biologically relevant and reasonable.

The driving forces of the dynamics for the metabolite concentration are described as below:

$$\dot{Y}_i = F(Y_i) = \sum_{j=1}^{N_i} X_j r_j \tag{17}$$

In Eq. (17), the variable Y represents the metabolite concentration and F represents the driving force of Y. The force is the summation of enzyme kinetic velocity r_j multiplied by the related enzymes X_j.

In real dynamics, fluctuations cannot be ignored. When including these effects, noises should be added to the above deterministic equations. The steady state probability landscape and the corresponding probability flux can be obtained by either the self-consistent mean field approach or by the Langevin simulations [79, 85].

The researchers obtained the landscape of cancer metabolism, using the self-consistent mean field approximation. The landscape U is defined as $U = -ln\,(P\text{ss})$, which is directly related to the steady state probability distribution P_{ss}.

There are 53 variables(13 genes, 17 enzymes and 23 metabolites, with total of 53 nodes) in the cancer gene-metabolism integrative network, so there are 53 dynamical equations. The researchers choose two dimensions (LDH and PDH) for display, as it is impossible to visualize the landscape in 53-dimensions. Lactate dehydrogenase (LDH) is a key enzyme for switching away from TCA cycle and can reflect aerobic glycolysis flux. Pyruvate dehydrogenase (PDH) is the first enzyme component of pyruvate dehydrogenase complex (PDC), which contributes to transforming pyruvate into mitochondria for subsequential TCA cycle and oxidative phosphorylation. Four steady state attractors emerge on the landscape which are the normal state (N), the cancer OXPHOS state (P), the cancer glycolysis state (G) and the cancer intermediate state (I) attractors, as shown in Figure 13 A and B. It is obvious

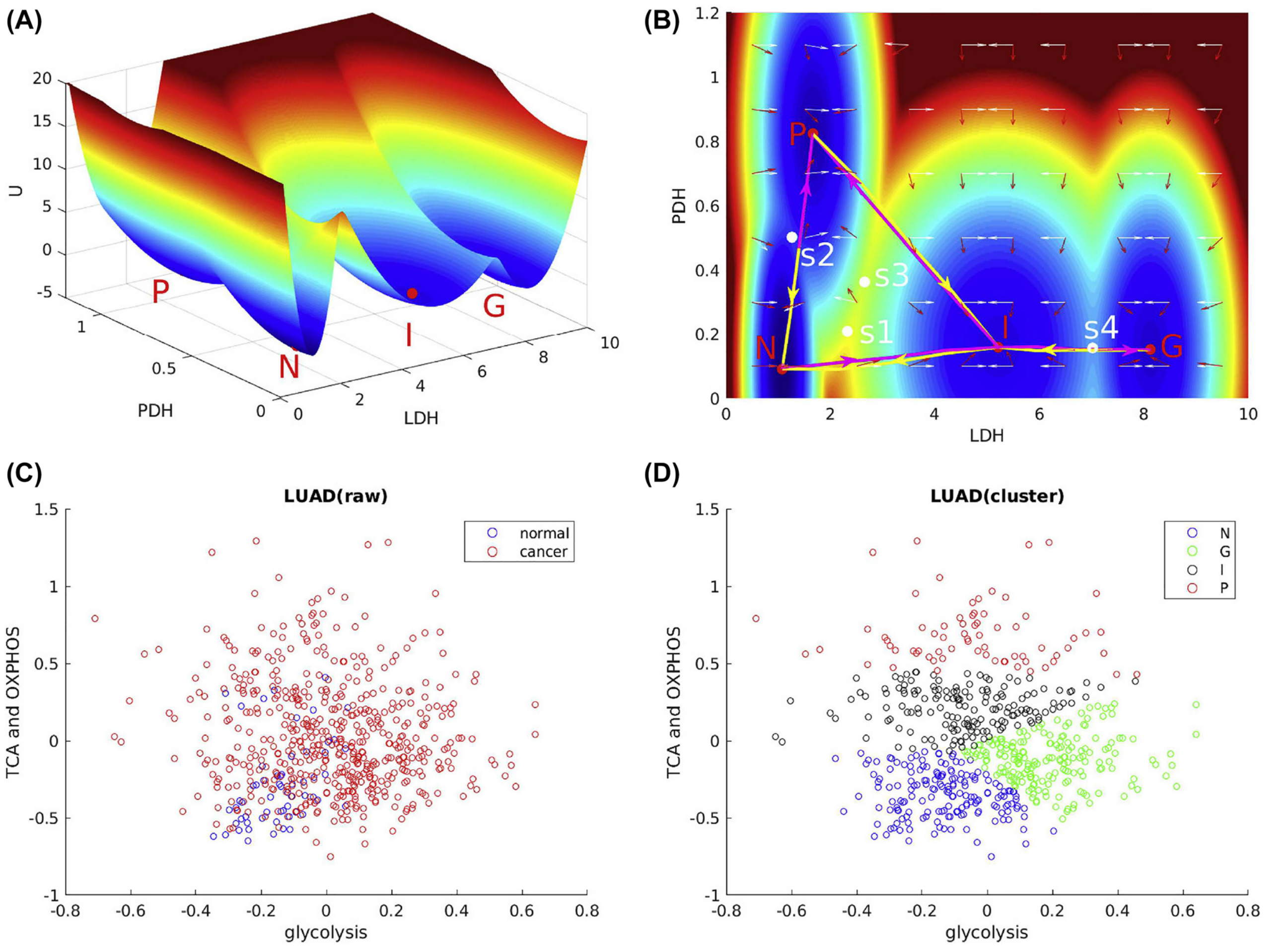

Figure 13: Landscape of cancer gene-metabolism and related gene expressions from GDC. N, P, G and I are normal state, cancer OXPHOS state, cancer glycolysis state and cancer intermediate state, respectively. s1, s2, s3 and s4 are saddle between normal state and cancer intermediate state, saddle between normal state and cancer OXPHOS state, saddle between cancer intermediate state and cancer OXPHOS state and saddle between cancer intermediate state and cancer glycolysis state, respectively. The yellow arrows represent the paths from N to I, from I to P, from N to P and from I to G; the magenta arrows represent the paths from I to N, from P to N, from P to I and from G to I. The white arrows represent the directions of the steady state probability flux, and the red arrow represent the directions of the negative gradient of the potential landscape. (A) The landscape of cancer gene-metabolism in 3D. (B) The landscape of cancer gene-metabolism in 2D. (C) Gene expression data with normal and cancer samples. (D) Gene expression data clustered by K-means (Image source: Li et al. [78]). PDH, pyruvate dehydrogenase; LDH, lactate dehydrogenase; LUAD, lung adenocarcinoma; TCA, tricarboxylic acid; OXPHOS, oxidative phosphorylation.

that the LDH/PDH level of the cancer intermediate state is lower compared to either cancer glycolysis state or cancer OXPHOS state. The red region represents the high potential area, while the blue region represents the low potential area. Between the two steady state attractors, there is a saddle which is colored white in Figure 13B. The researchers defined the saddle between the normal state and the cancer intermediate state as s1, the saddle between the normal state and the cancer OXPHOS state as s2, the saddle between the cancer intermediate state and the cancer OXPHOS state as s3, and the saddle between the cancer intermediate state and the cancer glycolysis state as s4.

In different tissues, cancer cells show different metabolic features [86]. As shown in Figure 13 (A and B), cells in the normal state consume lower ATP than the cells in the cancer state. The expression levels of both LDH and PDH are low. The expression level of PDH in the cancer OXPHOS state is much higher than that of the normal state. This is mainly related to the oxidative phosphorylation produced by the ATP and some cancer types such as melanoma and glioblastomata are oxidative phosphorylation-dependent [87]. On the other hand, the expression level of LDH in the cancer glycolysis state is much higher than that of the normal state. This is mainly related to the glycolysis produced by ATP and some cancer types such as liver and colorectal cancers are glycolysis-dependent [88]. The cancer intermediate state has lower PDH expression level than that of the cancer OXPHOS state and lower LDH expression level than that of the cancer glycolysis state. This state may correspond to the mixed cancer phenotype such as prostate cancer [89]. The cancer intermediate state can be seen as the bridge of the normal, the cancer OXPHOS and the cancer glycolysis state. The normal, OXPHOS and glycolysis states can therefore transform to each other through the cancer intermediate state.

The results from the model have been observed in the experiments in several cancer types. For example, the researchers used the RNA-seq data of lung adenocarcinoma (LUAD) from Genomic Data Commons Data Portal (GDC). The researchers normalized and averaged the gene expressions in each group. In Figure 13C, it can be seen that the glycolysis and OXPHOS levels of normal cells are much lower than that of cancer cells. This corresponds to the normal state (N) along with the cancer state(G, I, P) in the results of the model. The researchers further cluster these expressions into four groups as shown in Figure 13D. The four groups are consistent with the four states(N, G, I, P) of the model. These trends and results can also been observed in other cancer types.

Through the global sensitivity analysis of the underlying landscape topography, the researchers identified the key regulations which can promote cancer OXPHOS and glycolysis state. Moreover, the normal state to cancer state transformation or the bifurcations can be observed which is related to the peaks of the probability flux and the entropy production rate [77]. This can give a physical origin and a quantitative indicator of the cancer formation. In the model, the underlying dynamical and thermodynamical mechanism of cancer metabolism oscillations originated from the rotational steady state probability flux is uncovered. Based on the model, they provide some effectiveness of various metabolic therapeutic targets.

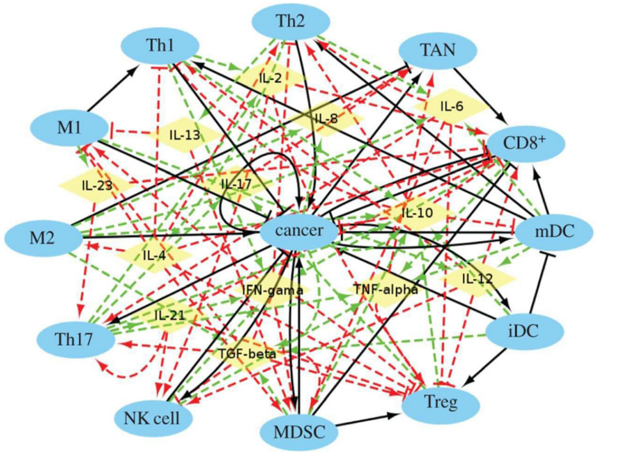

Figure 14: The cancer–immune network which includes 26 nodes and 107 regulations. The ellipses represent cells and yellow diamonds represent cytokines. Black solid arrows and bars represent cell-cell activation and inhibition, respectively. Red dashed arrows and bars represent cytokine–cell activation and inhibition, respectively. Green dashed arrows are cell–cytokine productions (Image source: Li et al. [90] with permission). MDSC, myeloid-derived suppressor cell; mDC, myeloid dendritic cells.

In the study [90], the researchers used landscape model to study the underlying mechanisms of the relationship between cancer and the immune system in tumorigenesis and cancer development. The researchers identified a network with cancer cells, 12 types of immune cells and 13 types of cytokines which facilitate the cell-cell communication. Figure 14 shows that the network includes cell-cell interactions, cytokine-cell interactions as well as cell-cytokine production. Due to the complexity of the network, in order to view the whole map more clearly, only the cell-cell interaction are shown in Figure 15. Based on the network, the driving forces of the dynamics for the cell or cytokine concentrations can be described and the corresponding landscape was shown in Figure 16.

The researchers analyzed the immune network quantitatively. Three steady state basins emerge based on the landscape topography which are normal (N), low cancer (L) and high cancer (H) state. The landscape model can reveal the cancer development and evolution process as the landscapes present different characteristics of the three phases (elimination, equilibrium and escape) in immunoediting [91]. The researchers also quantified the origin of cancer immune oscillations and predicted three types

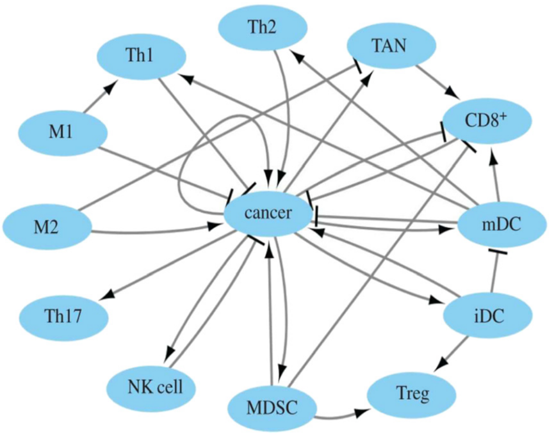

Figure 15: The network of the cancer immune system for cell-cell interactions. Black arrows and bars represent cell-cell activation and cell-cell inhibition, respectively (Image source: Li et al. [90] with permission). MDSC, myeloid-derived suppressor cell; mDC, myeloid dendritic cells.

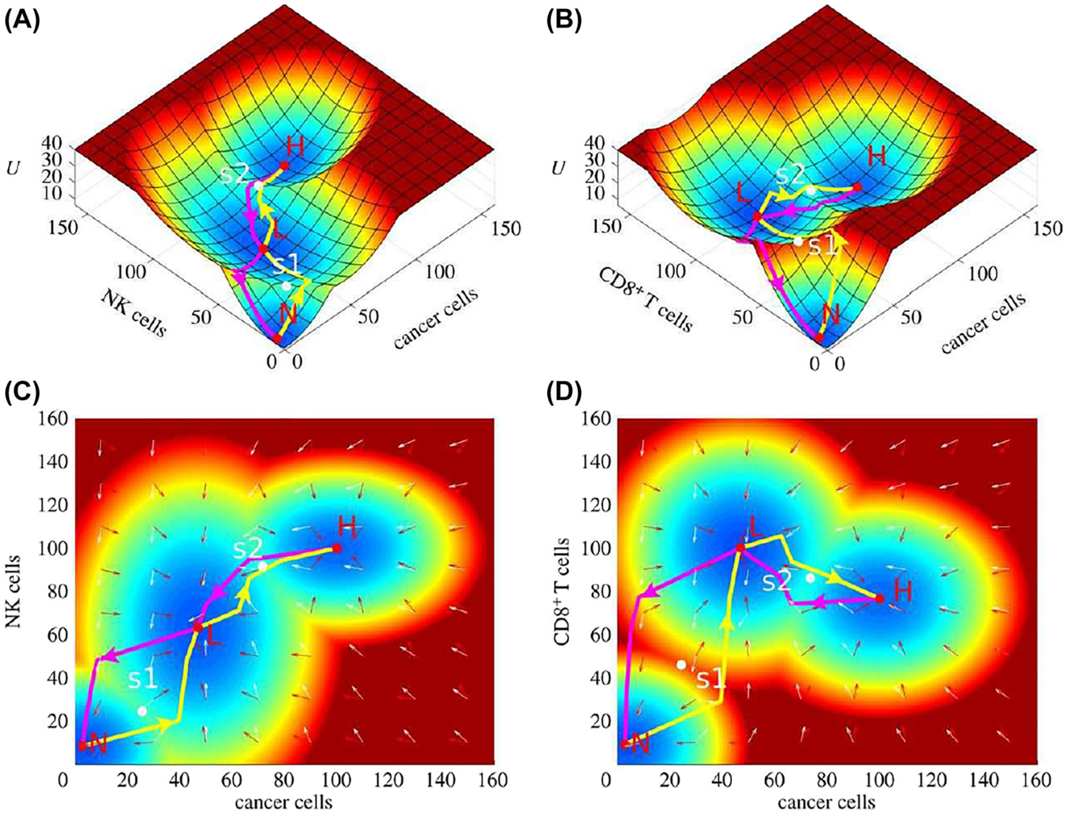

Figure 16: The landscape of cancer innate immune (A and C) which is depicted by cancer cells and natural killer cells. The landscape of adaptive immune system (B and D) which is depicted by cancer cells and CD8$^+$ T cells. N, L and H represent normal state, low cancer state and high cancer state, respectively. s1, s2 represent the saddle between normal state and low cancer state, saddle between normal state and high cancer state, respectively (Image source: Li et al. [90] with permission).

of immune cells (mDCs, NK cells and CD8$^+$) and two types of cytokines (IL-10 and IL-12) which are important immunotherapy targets.

Comparison of the ordinary differential equation and the stochastic differential equation models

Both the ordinary differential equation and the stochastic differential equation models are suitable for analyzing the non-linear systems and network dynamics. The two models have similarities and differences.

First, both approaches are based on the gene regulatory network through experimental data and the gene networks are all with directions and feed-back loops. The extensive experimental literature can provide important references for the selection of parameters and kinetic laws. The gene regulatory network construction by text mining can make full use of computing resources and establish a relative large and reliable network.

Second, both approaches used the ODEs to describe the variations of the gene expression levels in time. Moreover, the Hill function can be used in the equations to represent the activation or repression regulations which can describe the network dynamics in a quantitative way. In cellular environment, intrinsic and external fluctuations cannot be ignored. The fluctuations were taken into account in the stochastic dynamical models.

Third, both approaches can give the stable steady states of the system. The ODE models can give deterministic trajectory and the fix points (steady state points and saddle points). The biological states can be defined as stable states by the steady state fixed points. As shown in Figure 6A, the stable states are represented by the steady state points. This can help to perform analysis on the stability of different regions (on the stable states). When the signals of the system change, the corresponding steady states will shift accordingly (Figure 6B). It is sometimes difficult to associate different steady states to one biological function. Researchers integrated stochastic methods with RACIPE named sRACIPE to achieve the stochastic analysis [67].

The SDEs models can capture the effects of both the intrinsic and extrinsic noises in biological process through the stochastic trajectory. In the SDEs, the term η can be used to simulate the noise. When the noise is very small, the steady states almost coincide with the ODE models. Although the individual stochastic trajectories are unpredictable, they can be described by the probability evolution which satisfies linear equation and is predictable. The resulting (steady state) probability distribution and the associated potential landscape ($U = -\ln P$) can characterize the global stability of the system (As shown in Figures 7, 10 and 15). We give an example to demonstrate the differences between ODEs and SDEs model in Figure 17. Figure 17A shows a gene regulatory network, then we can use ODEs to obtain Figure 17B starting from different initial values or use SDEs to obtain Figure 17C run after sufficient time. Both Figure 17B and C have two steady state points or stable states. In Figure 17B we can see that the steady state points

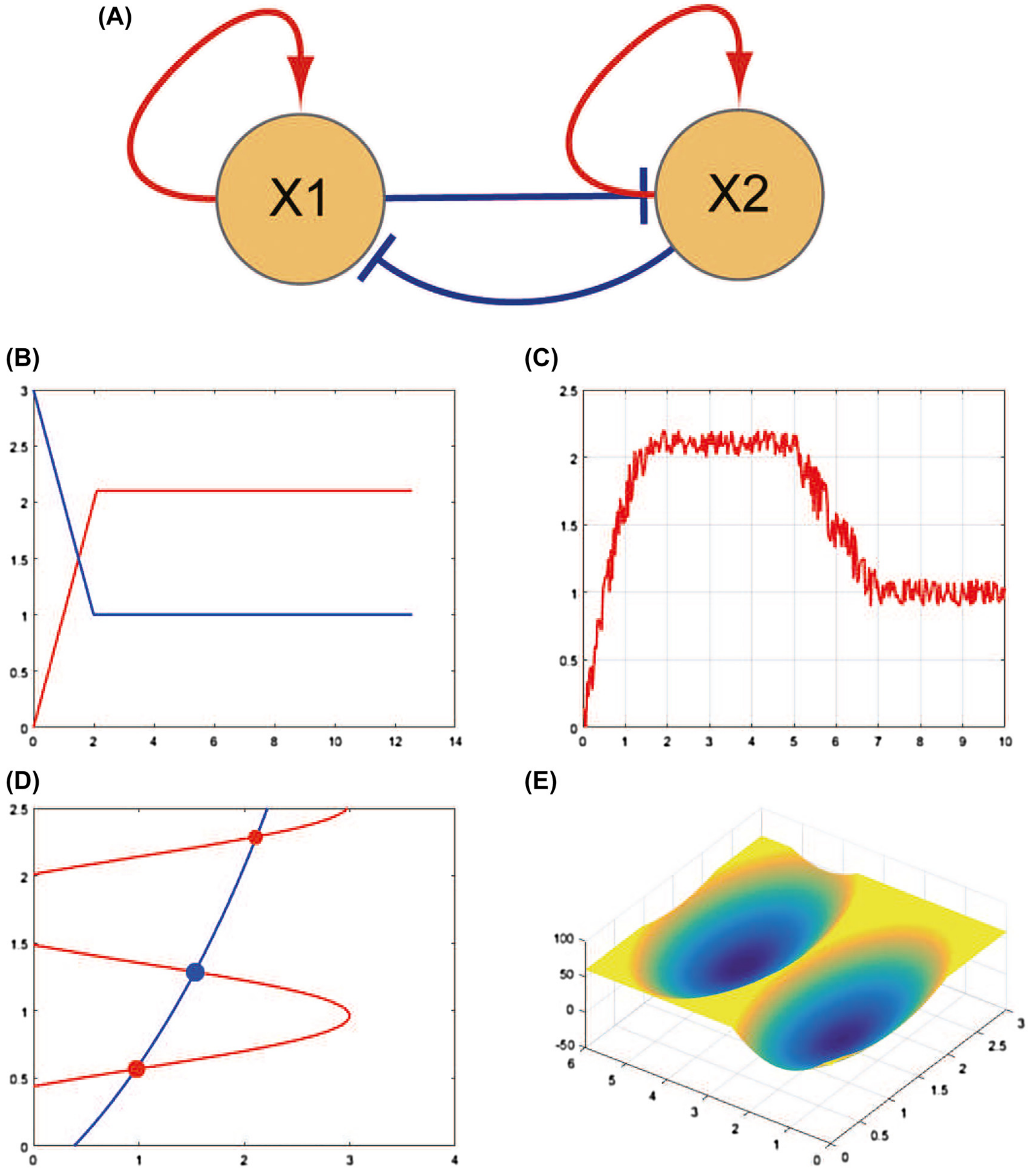

Figure 17: The comparisons of ODEs and SDEs models.
(A) is a gene regulatory network. (B) is obtained by ODEs which starting from different initial values. (C) is obtained by SDEs and run after suffificient time. (D) shows the deterministic trajectories and fixed points, where the red balls represent the steady state points and the blue ball represents the saddle point. (E) shows the landscape of the state space.

are 1 and 2.1, and in Figure 17C, the steady states are around the two values with the fluctuations. When we set the ODEs to be zero, we can find the fix points of the ODEs. In Figure 17D, the red balls represent the steady state points and the blue ball represents the saddle point. In theory, we can start from an initial value to the SDEs with certain noise and run after sufficient time to quantified the landscape as Figure 17E. In other words, while the ODEs give the deterministic trajectories and fixed points, the SDEs give the weight of the state or the landscape of the state space. For example, in the long time, ODEs give the fixed points with probability being equal to one while other point with zero probability. For SDEs, one fixed point has relatively higher probability, but other points in state space also have probability. It is in this sense that SDEs can give a global description of the state changes since each state is assigned to a different weight or probability. Be chose of this, the states can be connected and the description is global. This is the contrast with the ODEs case that there is no apparent connections between fixed points and the stability analysis is all local. Both the landscape and flux provide the driving force for the network dynamics. The landscape provides a global view of each functional states rather than limiting to the local changes such as gene mutations [92, 93]. Barrier heights and transition speed can be quantified to characterize the stability of each states [85]. The higher barrier heights can be used to infer the more stable state attractor and more difficulty to escape. The dominant paths can be used to quantified the actual process of state switching [94]. The regulations can influence the landscape topography, and the landscape topography can determine the stability of each state. Therefore, the global sensitivity analysis can help to find out key genes and key regulations which most likely influence the landscape topography. The landscape models can provide visualization and global quantification for analyzing different gene regulatory network at both the genetic and epigenetic levels.

Discussion

In this review, we reviewed several methods of machine learning which are widely used on cancer prediction/prognosis. The advantages and disadvantages of certain classical methods are summarized and related applications are given to help beginners to understand the application scenarios and the basic methods of the models. The core functions of the models are also introduced in this review which can help to understand the key ideas of the algorithm and its development in related fields. When one aims to predict cancer susceptibility, recurrence, survival etc., the data mining and machine learning algorithms can be applied to large data analysis. When and which algorithm should be chose is determined by the nature and quality of data, the specific problem for analysis and the kind of the output one aims. We reviewed some applications on cancer studies which focused on the development of predictive models. These models can help on cancer data classification and prediction. Most of the studies applied more than one algorithm to improve the prediction accuracy. Adding *a priori* information can also help to improve the predictive accuracy of the models. The data mining and machine learning methods have become a useful tool on diagnosis and treatment.

Then we discussed several computational approaches which are commonly used on gene regulatory networks identifications. Table 2 compared the advantages and disadvantages for each mentioned approach. When one aims to build an undirected gene regulatory network, one can use the Correlation method. This method can process large scale dataset as the low consumption of computing resources. When one aims to build a gene regulatory network with regulatory directions and enable to detect linear and non-linear interactions, one can use regression methods. This methods can also process large scale dataset. Simple Bayesian methods can build gene regulatory network with regulatory directions and handle logical interaction components with a small number of variables. When one used simple Bayesian methods, the prior knowledge are often integrated to strengthen the relationship between nodes. When one aims to build a gene regulatory network with with regulatory directions and feed-back loops, one can use the dynamical Bayesian methods. The input data of the dynamical Bayesian methods should be time series data. If there are missing values in the data, this will restrict the performance of dynamical Bayesian. Therefor, the data preprocessing, such as noise reduction, is very important for dynamical Bayesian method. The networks constructed by dynamical Bayesian are often small size networks.

At last, we introduced dynamics methods through ODEs and SDEs to study cancer using gene regulatory network. Both the two methods can be used to describe the network dynamics. The ODE methods can be used to analyze the stability of different regions by calculating the deterministic trajectory and the steady states of the system. Unlike the ODEs, SDEs introduce an additional stochastic force to describe the fluctuations of the system. While the SDEs methods can characterize the global stability of the system through the stochastic trajectories and the associated probability evolution of the system. These two methods are both suitable for analyzing the non-linear dynamics of gene regulatory networks.

There is no doubt that the data mining and machine learning technology can help to solve massive data analysis problems, it is still essential to discuss the limitations of these technologies. A major limitation of data mining and machine learning technology is that it is hard to explain how and why these algorithms make their conclusions. For example, ANNs can be seen as a black box which takes the inputs and produces outputs with no explanation of how it comes to the conclusions. Moreover, the accuracy of the results depends on the quality of the input data. 'Garbage in, garbage out.' is a very vivid illustration of the problem. If the input data are very poor, the output predictions will be inaccurate. DBNs' input data often need to have time series. As the experimental data are often pseudo time series data. Since the real time data are some times challenge to obtain (Gene expression has a high dimensional data structure, so dimensionality reduction is needed in the segmentation process. Time series are usually defined in terms of experimental intervals or artificially. These time series may not have logical links between the time points due to nature RNA-seq, for example, the cells at one time point measurement will not continue to be measured in the next time point. After dimensionality reduction and artificial time order, the data becomes pseudo time serials data), it is also crucial to consider the bias in the algorithms which are used to process these data. This limitation reinforces the prediction results of the algorithms. The data mining and machine learning can be a useful tool to supplement diagnosis and treatment, but it should not replace decisions based on clinical evidence. On the other hand, the dynamics methods used ODEs or SDEs to describe the biological process, the 'black-box' can be avoided. However, a simple biological process often requires many dynamical variables and associated equations to describe. Therefore, the analyzed systems are often very small or only limited for a motif. In recent years, this problem has been improved with the rise of parallel computing, but it will still take efforts to solve practical medical problems accurately and completely.

The success of data mining and machine learning technology in practice not only depend on the development of the new methods but also crucially rely on the experts or researchers to preprocess data, select or construct appropriate features or models, and evaluate the model with regard to the generalization, model accuracy, and risk profiles. The complexity of those mentioned can be challenging which we need to make progress in the future.

Research funding: CY was supported by National Nature Science Foundation of China Grants No. 21721003.

Competing interests: There is no conflict of interest in this manuscript.

References

1. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk factors and preventions of breast cancer. Int J Biol Sci 2017;13:1387–97.
2. Plummer M, de Martel C, Vignat J, Ferlay J, Bray F, Franceschi S. Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Global Health 2016;4:e609–16.
3. Sanmiguel P. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 1998;82:37–44.
4. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci Unit States Am 1950;36:344–55.
5. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (svm) learning in cancer genomics. Cancer Genomics Proteomics. 2018;15:41–51.
6. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8–16.
7. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. IEEE Access 2019;7:53040–65.
8. Kristensen VN, Christian Lingjærde O, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 2014;14: 299–313.
9. Fatima N, Liu L, Hong S, Ahmed H. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access 2020;8:150360–76.
10. Castaldo R, Cavaliere C, Soricelli A, Salvatore M, Pecchia L, Franzese M. Radiomic and genomic machine learning method performance for prostate cancer diagnosis: systematic literature review. J Med Internet Res 2021;23:e22394.
11. Madan Babu M. Evolution of transcription factors and the gene regulatory network in escherichia coli. Nucleic Acids Res 2003; 31:1234–44.
12. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Briefings Bioinf 2018;19: 325–40.
13. Jana S. Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. New Phytol 2019;228:35–41.
14. Nagalakshmi U, Wang Z, Karl W, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008;320:1344–9.
15. Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. Briefings Bioinf 2009;10:490–7.
16. Johnson DS, Ali M, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316:1497–502.
17. Potkin SG, Macciardi F, Guffanti G, Fallon JH, Wang Q, Turner JA, et al. Identifying gene regulatory networks in schizophrenia. Neuroimage 2010;53:839–47.

18. Wilczynski B, Furlong EEM. Challenges for modeling global gene regulatory networks during development: insights from drosophila. Dev Biol 2010;340:161–9.
19. Lee W-P, Tzou W-S. Computational methods for discovering gene networks from expression data. Briefings Bioinf 2009;10: 408–23.
20. Peter I, Schmalfuss B. J Dynam Differ Equ 2001;13:215–49.
21. Slavík A. Generalized differential equations: differentiability of solutions with respect to initial conditions and parameters. J Math Anal Appl 2013;402:261–74.
22. Justo-Silva R, Ferreira A, Flintsch G. Review on machine learning techniques for developing pavement performance prediction models. Sustainability 2021;13:5248.
23. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: a tutorial. Computer 1996;29:31–44.
24. Papadopoulos A, Fotiadis DI, Likas A. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines. Artif Intell Med 2005;34:141–50.
25. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited. Cancer 2010;116:3310–21.
26. Temkin NR, Holubkov R, Machamer JE, Richard Winn H, Dikmen SS. Classification and regression trees (CART) for prediction of function at 1 year following head trauma. J Neurosurg 1995;82:764–71.
27. Thomas G. Dietterich. Mach Learn 2000;40:139–57.
28. Alexander S, Bilchik A, Smith D, Eberhardt JS, Ben Ward E, Nissan A, et al. Clinical decision support and individualized prediction of survival in colon cancer: bayesian belief network model. Ann Surg Oncol 2012;20:161–74.
29. Mahadevan S, Ramesh R. Validation of reliability computational models using bayes networks. Reliab Eng Syst Saf 2005;87: 223–32.
30. Fatih Akay M. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009;36: 3240–7.
31. Waddell M, Page D, Shaughnessy J. Predicting cancer susceptibility from single-nucleotide polymorphism data. ACM Press; 2005;21–8.
32. Chen Y-C, Wan-Chi K, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med 2014;48:1–7.
33. Xu X, Zhang Y, Zou L, Wang M, Ao L. A gene signature for breast cancer prognosis using support vector machine. 2012 Int Conf Biomed Eng Inform 2012;928–31.
34. Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. Bioinformatics 2006;22:e184–90.
35. Rosado P, Lequerica-Fernández P, Villallaín L, Peña I, Sanchez-Lasheras F, de Vicente JC. Survival model in oral squamous cell carcinoma based on clinicopathological parameters, molecular markers and support vector machines. Expert Syst Appl 2013;40: 4770–6.
36. Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Eng Appl Artif Intell 2013;26:2194–205.
37. Exarchos KP, Goletsis Y, Fotiadis DI. Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans Inf Technol Biomed 2012;16:1127–34.
38. Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics 2006;23:30–7.
39. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109.
40. Urbanowicz RJ, Angeline SA, Rita Karagas M, Moore JH. Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach. J Am Med Inf Assoc 2013;20:603–12.
41. Kim W, Kim KS, Lee JE, Noh D-Y, Kim S-W, Jung YS, et al. Development of novel breast cancer recurrence prediction model using support vector machine. J Breast Cancer 2012;15:230.
42. Stojadinovic M, Stojadinovic M, Pantic D. Decision tree analysis for prostate cancer prediction. Srp Arh Celok Lek 2019; 147:52–8.
43. Delen D, Walker G, Amit K. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113–27.
44. Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Adrian D, et al. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res 2004;10:2725–37.
45. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999;21:33–7.
46. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.
47. Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr Opin Biotechnol 2019;58:129–36.
48. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4: Article17.
49. Xu P, Yang J, Liu J, Yang X, Liao J, Yuan F, et al. Identification of glioblastoma gene prognosis modules based on weighted gene co-expression network analysis. BMC Med Genom 2018;11:96.
50. Tian F, Zhao J, Fan X, Kang Z. Weighted gene co-expression network analysis in identification of metastasis-related genes of lung squamous cell carcinoma based on the cancer genome atlas database. J Thorac Dis 2017;9:42–53.
51. Jiang H, Huang Q, Chen L, Li Z, Xu Y, Sun H, et al. Multi-classification of cancer samples based on co-expression analyses. 2019 IEEE Int Conf Bioinform Biomed 2019;197–201.
52. Wu Y, liu F, Luo S, Yin X, He D, Liu J, et al. Co-expression of key gene modules and pathways of human breast cancer cell lines. Biosci Rep 2019;39:BSR20181925.
53. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat 2004;32:407–51.
54. Werhli AV, Husmeier D. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. Stat Appl Genet Mol Biol 2007;6:Article15.
55. Kung Tan A, Saberi Mohamad M. Using bayesian networks to construct gene regulatory networks from microarray data. Jurnal Teknologi 2012;58:1–6.
56. Care MA, Westhead DR, Tooze RM. Parsimonious gene correlation network analysis (pgcna): a tool to define modular gene co-expression for refined molecular stratification in cancer. NPJ Syst Biol Appl 2019;5:13.
57. Haury AC, Mordelet F, Vera-Licona P, Vert JP. TIGRESS: trustful inference of gene REgulation using stability selection. BMC Syst Biol 2012;6:145.

58. Friedman N, Linial M, Nachman I, er DP. Using bayesian networks to analyze expression data. J Comput Biol 2000;7:601–20.
59. Adabor ES, Acquaah-Mensah GK. Restricted-derestricted dynamic bayesian network inference of transcriptional regulatory relationships among genes in cancer. Comput Biol Chem 2019;79:155–64.
60. Dojer N, Anna G, Mizera A, Wilczyński B, Tiuryn J. BMC Bioinf 2006;7:249.
61. Nguyen X, Chetty M, Ross C, Wangikar PP. Gene regulatory network modeling via global optimization of high-order dynamic bayesian network. BMC Bioinf 2012;13:131.
62. Jia D, Lu M, Jung KH, Park JH, Yu L, Onuchic JN, et al. Elucidating cancer metabolic plasticity by coupling gene regulation with metabolic pathways. Proc Natl Acad Sci Unit States Am 2019;116: 3909–18.
63. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, et al. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. npj Digital Med 2019;2:115.
64. Philipsen KR, Christiansen LE, Hasman H, Madsen H. Modelling conjugation with stochastic differential equations. J Theor Biol 2010;263:134–42.
65. Yu L, Lu M, Jia D, Ma J, Ben-Jacob E, Levine H, et al. Modeling the genetic regulation of cancer metabolism: interplay between glycolysis and oxidative phosphorylation. Cancer Res 2017;77: 1564–74.
66. Lu M, Huang B, Hanash SM, Onuchic JN, Ben-Jacob E. Modeling putative therapeutic implications of exosome exchange between tumor and immune cells. Proc Natl Acad Sci Unit States Am 2014; 111:E4165–74.
67. Kohar V, Lu M. Role of noise and parametric variation in the dynamics of gene regulatory circuits. npj Syst Biol Appl 2018;4:40.
68. Chalancon G, Ravarani CNJ, Balaji S, Martinez-Arias A, Aravind L, Jothi R, et al. Interplay between gene expression noise and regulatory network architecture. Trends Genet 2012;28:221–32.
69. Cole JA, Luthey-Schulten Z. Careful accounting of extrinsic noise in protein expression reveals correlations among its sources. Phys Rev 2017;95:062418.
70. Tkačik G, Gregor T, Bialek W. The role of input noise in transcriptional regulation. PLoS One 2008;3:e2774.
71. Goychuk I, Jung P, Kohler S, Schmid G, Talkner P. Stochastic processes in physics and chemistry (in honor of peter hänggi). Chem Phys 2010;375:131–2.
72. Sasai M, Wolynes PG. Stochastic gene expression as a many-body problem. Proc Natl Acad Sci Unit States Am 2003;100: 2374–9.
73. Zhang K, Sasai M, Wang J. Eddy current and coupled landscapes for nonadiabatic and nonequilibrium complex system dynamics. Proc Natl Acad Sci Unit States Am 2013;110:14930–5.
74. Wang J, Xu L, Wang EK. Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation, and coherence of biochemical oscillations. Proc Natl Acad Sci USA 2008;105:12271–6.
75. Wang J. Landscape and flux theory of non-equilibrium dynamical systems with application to biology. Adv Phys 2015; 64:1–137.
76. Wang J, Zhang K, Xu L, Wang E. Quantifying the waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci Unit States Am 2011;108: 8257–62.
77. Qian H. Mesoscopic nonequilibrium thermodynamics of single macromolecules and dynamic entropy-energy compensation. Phys Rev 2001;65:016102.
78. Li W, Wang J. Uncovering the underlying mechanisms of cancer metabolism through the landscapes and probability flux quantifications. iScience 2020;23:101002.
79. Chong Y, Wang J. A physical mechanism and global quantification of breast cancer. PLoS One 2016;11:e0157422.
80. Li C, Wang J. Quantifying the landscape for development and cancer from a core cancer stem cell circuit. Cancer Res 2015;75: 2607–18.
81. Li C, Wang J. Quantifying the underlying landscape and paths of cancer. J R Soc Interface 2014;11:20140774.
82. Chong Y, Xu H, Wang J. A global and physical mechanism of gastric cancer formation and progression. J Theor Biol 2021;520: 110643.
83. Leung WK, Sung JJY. Intestinal metaplasia and gastric carcinogenesis. Aliment Pharmacol Ther 2002;16:1209–16.
84. Magalhaes PP. CagA status of helicobacter pylori infection and p53 gene mutations in gastric adenocarcinoma. Carcinogenesis 2003;24:145.
85. Wang J, Zhang K, Wang E. Kinetic paths, time scale, and underlying landscapes: a path integral framework to study global natures of nonequilibrium systems and networks. J Chem Phys 2010;133:125103.
86. Lehuédé C, Dupuy F, Rabinovitch R, Jones RG, Siegel PM. Metabolic plasticity as a determinant of tumor growth and metastasis. Cancer Res 2016;76:5201–8.
87. Obre E, Rossignol R. Emerging concepts in bioenergetics and cancer research: metabolic flexibility, coupling, symbiosis, switch, oxidative tumors, metabolic remodeling, signaling and bioenergetic therapy. Int J Biochem Cell Biol 2015;59: 167–81.
88. Graziano F, Ruzzo A, Giacomini E, Ricciardi T, Aprile G, Loupakis F, et al. Glycolysis gene expression analysis and selective metabolic advantage in the clinical progression of colorectal cancer. Pharmacogenomics J 2016;17:258–64.
89. Elia I, Schmieder R, ChristenS, Fendt SM. Organ-Specific Cancer Metabolism and Its Potential for Therapy. Handb Exp Pharmacol. 2016;233:321–53.
90. Li W, Wang J. Correction to 'uncovering the underlying mechanism of cancer tumorigenesis and development under an immune microenvironment from global quantification of the landscape'. J R Soc Interface 2021;18:20210247.
91. Dunn GP, Bruce AT, Ikeda H, Lloyd JO, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol 2002;3:991–8.
92. Chong Y, Liu Q, Chen C, Wang J. Quantification of the underlying mechanisms and relationships among cancer, metastasis, and differentiation and development. Front Genet 2020;10:1388.
93. Xu L, Zhang K, Wang J. Exploring the mechanisms of differentiation, dedifferentiation, reprogramming and transdifferentiation. PLoS One 2014;9:e105216.
94. Li C, Wang J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. PLoS Comput Biol 2013;9:e1003165.