GMR
Genetics and Molecular Research
ISSN 16765680

Time-series microarray data simulation modeled with a case-control label

Y.J. Liu and J.Y. Zhang

School of Computer Science and Technology, Xidian University, Xi’an, China

Corresponding author: J.Y. Zhang
E-mail: jyzhang@mail.xidian.edu.cn

Genet. Mol. Res. 15 (2): gmr.15027287
Received August 22, 2015
Accepted January 14, 2016
Published May 13, 2016
DOI http://dx.doi.org/10.4238/gmr.15027287

ABSTRACT. With advances in molecular biology, microarray data have become an important resource in the exploration of complex human diseases. Although gene chip technology continues to grow, there are still many barriers to overcome, such as high costs, small sample sizes, complex procedures, poor repeatability, and the dependence on data analysis methods. To avoid these problems, simulation data have a vital role in the study of complex diseases. A simulation method of microarray data is introduced in this study to model the occurrence and development of general diseases. Using classic statistics and control theory, five risk models are proposed. One or more models can be introduced into the baseline simulation dataset with a case-control label. In addition, time-series gene expression data can be generated to model the dynamic evolutionary process of a disease. The prevalence of each model is estimated and disease-associated genes are tested by significance analysis of microarrays. The source code, written in MATLAB, is freely and publicly available at http://sourceforge.net/projects/genesimulation/files/.

Key words: Microarray data simulation; Risk model; Case-control label; Time-series data

INTRODUCTION

Normal gene expression has space-time specificity. Conversely, abnormal gene expression is often considered as the root of disease, e.g., oncogene activation or tumor suppressor gene inactivation in cancer. Most genes are regulated during the transcription process, and gene expression values are often considered specific physiological or biological disease status markers. Additionally, the development of a complex disease is a dynamic process with individual gene specificities, and is the result of additive genetic mutations. In the post genome era, time-series gene expression data have become an important resource to explore complex human diseases.

Although gene chip technology continues to advance, there are still many barriers to overcome, such as high costs, small sample sizes, complex procedures, poor repeatability, and dependence on data analysis methods. To avoid these problems, simulation data have become a valuable resource for complex disease studies (Carvajal-Rodríguez, 2008; Hoban et al., 2012).

Various methods for simulating static microarray data representing only one time point have been developed. Singhal et al. (2003) provided a microarray data simulator, which modeled “normal tissue samples” and “diseased tissue samples” with known defined changes in gene expression, where the changes were estimated by analyzing microarray hybridization experiments. Albers et al. (2006) developed a simulator called SIMAGE, which focused on multiple layers of factors influencing microarray experiments, guaranteeing that the simulated data mimicked real data as closely as possible. Nykter et al. (2006) presented a microarray simulation model, which included all of the steps that affect the quality of real microarray data. This model also included the simulation of biological ground truth data, application of biological and measurement technology specific error models, and finally simulation of the microarray slide manufacturing and hybridization process.

In regards to time-series gene expression research, a great deal of focus has been placed on reverse-engineering or modeling gene regulatory networks. However, the prediction and simulation of time-series gene expression data have received little attention. So far, there are few specific disease deterioration models. Vu and Vohradsky (2002) designed a simulator for dynamics of genetic regulatory networks. The model was based on the recurrent neural network principle, and allowed for interactive simulation of various genetic regulatory interactions under different features of the system. However, the number of samples and genes were small and there was no disease-related information. Bresch et al. (2010) focused on tumor progression by using partial differential equations regarding tumor size and tumor cell number. Additionally, researchers from Miami and Heidelberg developed a mathematical model to predict tumor growth trends (Choe et al., 2011).

Gene expression varies with time and space, and mutual interactions exist between them. Even if the behavior of expression of a single gene may be simple, the collective expression behavior of multiple genes is complex. Hence, it may be impossible to find a completely generalizable model for complex biological networks. Based on classical statistics and control theory, we propose risk models of general diseases, and present a simulation method for microarray data. The resultant simulated time-series data with case-control tags model the occurrence and development of complex diseases.

MATERIAL AND METHODS

An overview of the simulation method

The proposed simulation model is modular. The structure of the model is presented in Figure 1, and is divided into the following parts:

Baseline microarray data simulation model

In real microarray experiments, gene expression data can be obtained under certain conditions. Similarly, the first step of simulation work is to simulate the start point (t_0) of gene expression data as baseline data, which constitutes a data matrix M_{N*G}. Each element in m_{ij} represents an expression value of the j^{th} gene in the i^{th} sample, the row vector $M_{i.} = (m_{i1}, m_{i2}, …, m_{iG})$ represents the expression level of each gene from a sample, and the column vector $M_{.i} = (m_{1i}, m_{2i}, ..., m_{Ni})$ represents the expression level of gene i in various samples.

Biological and measurement noise model

In real microarray experiments, preprocessing is important prior to analyzing the obtained data, which includes data cleaning, filtering, and transformation. On the contrary, in simulation experiments, it is necessary to add noise data to mimic biological noise and measurement errors.

Disease occurrence model

There have been numerous simulator studies on individual phenotypic information (sick or well) with genome-wide single nucleotide polymorphism (SNP) data (Yuan et al., 2012). In reference to SNP studies, risk models based on gene expression data have been conducted using relevant statistics (Tang et al., 2009), which will be explained in more detail in the following section. According to these models, samples are marked with a case-control label.

Disease development model

As diseases develop, the expression of certain genes changes markedly. In general, the changes in the expressed genes account for a small portion of total gene expression, while most genes are either simply expressed or not. Therefore, a large number of flat curves are present in the gene expression profile. However, a small number of genes such as disease-associated genes or cell cycle-related genes have changes that are much bigger.

A cell is considered a complicated dynamic system, and each gene is equivalent to one variable of this system. In this system, linear and nonlinear development models are applied to simulate the process of disease development.

Baseline microarray data simulation model

The first step of the proposed simulation work is to simulate the start point (t_0) of gene expression data as baseline data. In real gene chip experiments, the log ratio of the two types of fluorescence intensity is used to determine gene expression. Assuming the absence of other systematic deviations, the log ratio is zero for non-regulated genes (with deviations only due to randomness), positive for up-regulated genes, and negative for down-regulated genes. In reference to the current simulator SIMAGE, the simulated log expression x_i of the j^{th} gene in the i^{th} sample is denoted as: $x_{ij} = G_{ij} + D_{ij}$ (Equation 1) (Albers et al., 2006), where G_{ij} is concerned with the 'true' expression of gene j in the i^{th} sample, and D_{ij} is a (possible) deviation due to up- or down-regulation of gene j in the i^{th} sample.

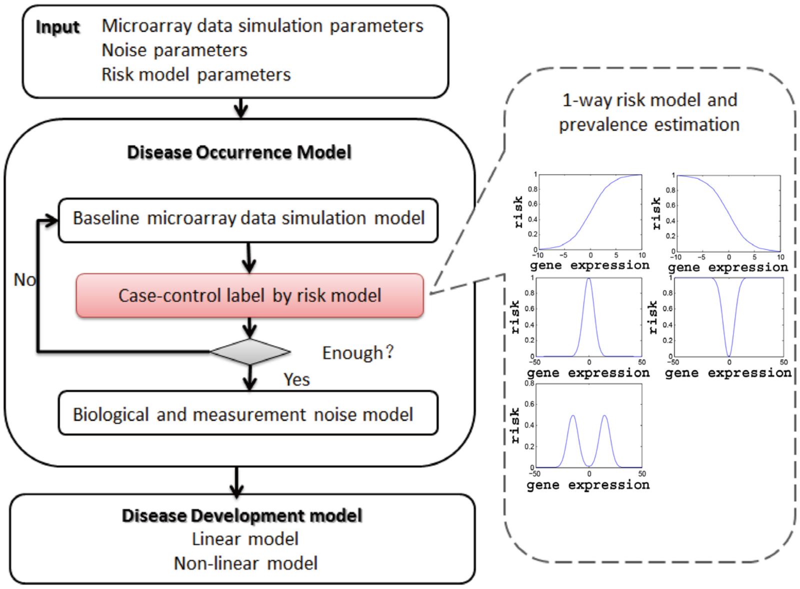

Input
Microarray data simulation parameters
Noise parameters
Risk model parameters
Disease Occurrence Model
Baseline microarray data simulation model
No
Case-control label by risk model
Enough?
Yes
Biological and measurement noise model
Disease Development model
Linear model
Non-linear model
1-way risk model and
prevalence estimation
risk
gene expression

Figure 1. Block diagram of the microarray simulation models.

To start with G_i, we assume that the 'true' expression level of gene *i* in various samples follows a normal distribution as $N(\mu, s^2_G)$. Here, μ is the expression mean, and s^2_G can be interpreted as the variation in 'true' expression. Hence, the expressions of non-differentially expressed genes are distributed symmetrically around μ, according to a normal distribution.

Additionally, the effect of regulation is modeled via D_i. For each sample, we model the probability of up- and down-regulated genes as λ_+ and λ_-, respectively, and the degree of up- or down-regulation of genes as a function of its baseline expression. A random number is generated as the degree of up- or down-regulation, and follows a uniform distribution along ($\mu_- \sim \mu_+$).

In summary, the ratio value is split into the aforementioned two parts, and the parameters are specified in Table 1.

Table 1. Overview of parameters in the baseline simulation model.

Parameter	Description
N	Number of genes
caseNum	Number of case samples
controlNum	Number of control samples
μ	Mean in 'true' expression
σ^2_G	Variance in 'true' expression
μ_+	Max change due to regulation
μ_-	Min change due to regulation
λ_+	Percent of up-regulated genes
λ_-	Percent of down-regulated genes

Biological and measurement noise model

Real microarray experiments are semi-quantitative analyses, which hence have some errors or disturbances. There have been numerous studies characterizing the properties of the error sources. Biological variations are typically considered to include the internal stochastic noise of the cells and error sources related to sample preparation. Experimental systematic errors include two categories: the error caused by the chip manufacture, and the error caused by the sample detection process (Nykter et al., 2006). Therefore, biological and measurement errors should be added to the simulated data, which are expressed as $m_{ij} = x_{ij} + e$ (Equation 2), where e is an error term, and m_{ij} is the observed expression value of gene j in the i^{th} sample. Depending on the user's purpose, the software offers two error models for users to choose from, as shown in Table 2.

Table 2. List of error models.

Simple EM:	
Model	Additive Gaussian noise is added to the data.
μ	Mean of the additive Gaussian noise. Noise is drawn from $N(\mu, \sigma^2)$
σ^2	Variance of the additive Gaussian noise.
SNR EM:	
Model	Additive Gaussian noise is added to the data with given signal-to-noise ratio.
μ	Mean of the additive Gaussian noise.
SNR	Signal-to-noise ratio after the noise is added.

Disease occurrence model

The most important part of the proposed simulation model herein is the disease occurrence model. In regards to risk model based on SNPs (Tang et al., 2009), the following risk models are used to model the phenotypic probability of genes as a cause of a disease. Different from SNP data, gene expression data are continuous rather than discrete. Additionally, there are a number of genes that can jointly affect a causal disease model, or jointly influence the probability of a resultant phenotype of a sample. So without loss of generalizability, one can simply assume gene indices of 1, 2, ..., r are disease-related, i.e., the gene set $G = \{1, 2, ..., r\}$ is disease-related with the genes jointly contributing to the probability of a phenotype in an r-way model.

We assume the disease-risk of a sample as the risk function $R(g)$, which determines the expression levels of the disease-related genes to be $g = (g_1, g_2, ..., g_r)^T$. The risk function $R(g)$ can generally be of multiple types across different models (in a multiple model situation) for the same disease case, and it can be so for models across different diseases due to the complexity of biologically mechanized disease scenarios and disease subtypes.

Based on the above concepts, we use the structures of sigmoid, Gaussian, and Gaussian mixture model methods, because these are flexible and rigorous methods, which are suitable for modeling various stochastic phenomena. In this study, we formulate five types of 1-way models, which are shown in Table 3. Comparing the risks calculated by these models with stochastic risk, the samples are marked with case-control labels.

Disease development model

Time-series microarray data can capture dynamic genomic behavior, which is not available in steady-state expression data such as that for disease progression and drug response. In advanced, multicellular living bodies, each gene may be affected by a specific subset of genes. The presence and expression patterns of most genes are to fulfill a special set of functions. Gene expression values can be used to build networks, in which associated actions and regulation of genes depend on the

Table 3. List of risk models.

We assume that there is only one disease gene g	
Model 1	$R(m) = 1 / (e^{-a(m-c)} + 1)$
DEscription	The risk function is a sigmoid function. The disease gene corresponds to an oncogene. Higher gene expression values indicate increased risk of pathogenesis.
PArameters	α is a tilt factor. When the gene expression value is at point *c*, the risk function *R*(*m*) is equal to 0.5.
Model 2	$R(m) = 1 / e^{a(m-c)} + 1$
DEscription	The risk function is an inverse sigmoid function. The disease gene corresponds to a tumor suppressor gene. Higher gene expression values indicate reduced risk of pathogenesis.
PArameters	α is a tilt factor. When the gene expression value is at point *c*, the risk function *R*(*m*) is equal to 0.5.
Model 3	$R(m) = N(\mu, \sigma^2)$
DEscription	The risk function is a Gaussian function. If the gene expression value is at point μ, there is the highest prevalence.
PArameters	μ is expectation, σ is variance.
Model 4	$R(m) = 1 - N(\mu, \sigma^2)$
DEscription	The risk function is an inverse Gaussian function. If the gene expression value is at point μ, there is the lowest prevalence.
PArameters	μ is expectation, σ is variance.
Model 5	$R(m) = \sum_{i=1}^{M} \alpha_i N_i(\mu_i, \sigma_i^2) \quad (\sum_{i=1}^{M} \alpha_i = 1)$
DEscription	The risk function has many peaks.
PArameters	i is the number of the mixed Gaussian functions, μ_i is expectation, σ_i is variance, α_i is weighting coefficient.

cell structure, gene function, and the environment. Therefore, it is possible to simulate the course of disease development in gene expression by regulatory gene networks. However, based on the large number of genes that dynamically influence each other, network building processes are still mathematically challenging as gene regulatory networks are complex, continuous, and dynamic. There are many current methods of building gene regulatory networks including Boolean networks, Bayesian networks, neural network models, and network model equations.

Among current methods, the differential equation model is the most accurate method to represent relationships in dynamic networks. There are many models that have been proposed based on differential equations, and ordinary differential equations methods are widely used (De Smet and Marchal, 2010; Kuwahara et al., 2013). Generally, the expression of one gene at one time point within a period is indicated by a function that describes its effects (i.e., activation or inhibition), and often involves the regulation of other genes. This relationship can be expressed by the following equation: $dm_{ij} / d_t = f\,(m_{i1}, m_{i2}, ..., m_{in})$ (Equation 3), where m_{ij} ($j \leq$ n) is an expression value of gene *j* in the i^{th} sample at time *t*, and function *f* may be linear or non-linear continuous.

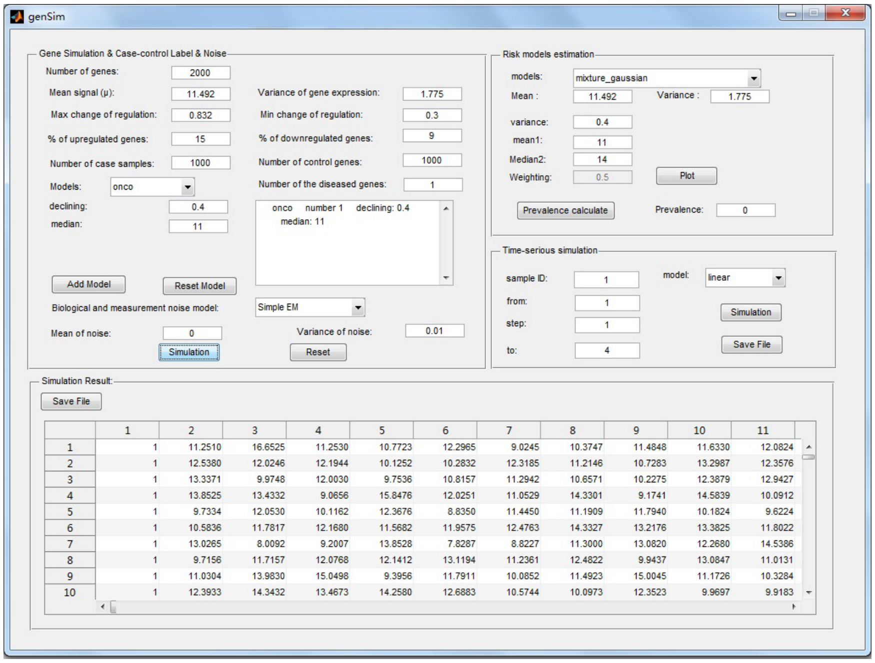
The linear analytical method is expressed as $dm_{ij} / d_t = \Sigma\ w_{jw}m_{iw} + b_{ij}$ from *w = 1 to w = n* (Equation 4), where w_{jw} represents the relation of the regulatory gene *w* to the target gene *j*, and b_{ij} represents the base activity. w_{jw} is positive, negative, or zero, which indicates activation, inhibition, or no obvious regulatory relationships, respectively.

Based on the current tool Genexp (Vu and Vohradsky, 2002), a sigmoid function is used as the nonlinear model expressed as: $dm_{ij} / d_t = k_{1j}$ [1 / (1 + exp(-$\Sigma\ w_{ij}m_{ij}$ - b_{ij}))] – (ln 2 / $t_{1/2j}$)m_{ij} (Equation 5), where m_{ij} is the expression of gene *j* at time point *t*, k_{1j} represents the cumulative ratio of gene *j*, and $t_{1/2j}$ represents the protein half-life of gene *j*.

RESULTS

Implementation

The tool is implemented entirely in MATLAB and works with MATLAB R2012b or a later version. Users specify parameters or choose default parameters through the input interface shown in Figure 2. Following the input phase, the simulation algorithm is executed and the results are shown in the result panel. Each row represents one sample. The first column is the disease label of the sample where 1 indicates a case sample, and 0 indicates a control sample. Each value in the other column represents the expression value of a gene.



Gene Simulation & Case-control Label & Noise

Number of genes: 2000
Mean signal (μ): 11.492
Variance of gene expression: 1.775
Max change of regulation: 0.832
Min change of regulation: 0.3
% of upregulated genes: 15
% of downregulated genes: 9
Number of case samples: 1000
Number of control genes: 1000
Models: onco
Number of the diseased genes: 1
declining: 0.4
median: 11

onco number 1 declining: 0.4
median: 11

Biological and measurement noise model: Simple EM
Mean of noise: 0
Variance of noise: 0.01

Risk models estimation

models: mixture_gaussian
Mean : 11.492
Variance : 1.775
variance: 0.4
mean1: 11
Median2: 14
Weighting: 0.5

Prevalence: 0

Time-serious simulation

sample ID: 1
model: linear
from: 1
step: 1
to: 4

Simulation Result:

	1	2	3	4	5	6	7	8	9	10	11
1	1	11.2510	16.6525	11.2530	10.7723	12.2965	9.0245	10.3747	11.4848	11.6330	12.0824
2	1	12.5380	12.0246	12.1944	10.1252	10.2832	12.3185	11.2146	10.7283	13.2987	12.3576
3	1	13.3371	9.9748	12.0030	9.7536	10.8157	11.2942	10.6571	10.2275	12.3879	12.9427
4	1	13.8525	13.4332	9.0656	15.8476	12.0251	11.0529	14.3301	9.1741	14.5839	10.0912
5	1	9.7334	12.0530	10.1162	12.3676	8.8350	11.4450	11.1909	11.7940	10.1824	9.6224
6	1	10.5836	11.7817	12.1680	11.5682	11.9575	12.4763	14.3327	13.2176	13.3825	11.8022
7	1	13.0265	8.0092	9.2007	13.8528	7.8287	8.8227	11.3000	13.0820	12.2680	14.5386
8	1	9.7156	11.7157	12.0768	12.1412	13.1194	11.2361	12.4822	9.9437	13.0847	11.0131
9	1	11.0304	13.9830	15.0498	9.3956	11.7911	10.0852	11.4923	15.0045	11.1726	10.3284
10	1	12.3933	14.3432	13.4673	14.2580	12.6883	10.5744	10.0973	12.3523	9.9697	9.9183

Figure 2. User interface. The image shows the interface of our simulator, which is divided into four sections. The top left section is the parameter set panel for baseline gene expression, risk model, and biological and measurement noise. In the top right section, a user estimates the risk model and prevalence. The middle right section is the time-series simulation plane. The bottom panel is the result panel.

Comparison of the five risk models

We assume the disease-risk of a sample is a risk function *R*(*g*), which is determined by the expression levels of the disease-associated genes. In our simulator, five typical types of 1-way models are presented for modeling various phenomena.

The risk function of the first model is a sigmoid function. The graph of a sigmoid function has a characteristic “S” shape. The disease gene corresponds to an oncogene, and higher gene expression values indicate higher probabilities of pathogenesis. Conversely, the second model’s risk function is an inverse sigmoid function. The disease gene corresponds to a tumor suppressor gene, where higher gene expression values indicate lower chances of pathogenesis.

The risk function of the third model is a Gaussian function. The graph of a Gaussian has a characteristically symmetric “bell curve” shape. μ is the position of the center of the curve peak. Therefore, if the gene expression value is at point μ, the risk is the highest. The risk gradually decreased on both sides of μ. Conversely, the fourth model risk function is an inverse Gaussian function. If the gene expression value is at point μ, the risk is the lowest, and the risk gradually increased on both sides of μ. The last model uses the mixed Gaussian function to mimic the situation when the risk curve has more than one peak. Graphical analysis is provided as a preliminary and intuitive judgment about the trend of the risk functions.

Prevalence estimation

Based on risk models of SNPs, risk models are used to model the phenotypic probability of genes as a cause of a disease. Here, we assume that gene indices of 1, 2, ..., r are disease-related, and gene set $G = \{1, 2, ..., r\}$ is disease-related, where genes jointly contribute to the probability of a phenotype in an r-way model.

It is simple to estimate the joint density function of these model-related genes from the gene expression dataset of $p(g) = p(g_1, g_2, ..., g_r)$, where $g = (g_1, g_2, ..., g_r)^T$ is an r-dimensional vector with its elements being the expression levels of the model-related gene set $G = \{1, 2, ..., r\}$. Therefore, based on SNP risk models, prevalence is estimated. The prevalence of the model $P(D)$ is dependent to the risk function $R(g)$ of the model, by an r-fold integral over the r-dimensional gene space defined by the gene indices of (1, 2, ..., r) expressed as $P(D) =$ ò$R(g)p(g)dg$ (6), where the risk function $R(g)$ determines the disease-risk of a sample, and the expression levels of the disease related genes are $g = (g_1, g_2, ..., g_r)^T$. The prevalence ranges from 0 to 1, where a higher value indicates a higher disease rate.

Comparison between the linear and non-linear methods

Time-series gene expression data have some biological features such as randomness, complexity, specificity, and dynamic characteristics, which are often difficult to simulate. Based on differential equations, the expression of one gene at one time point within a period is indicated by a function that describes its effects (i.e., activation or inhibition). In our study, linear and non-linear functions are provided.

The linear analytical method is less complex and easier to handle than the non-linear method. However, it is often used to solve simple problems, and in most cases, the interaction between genes exhibits complex nonlinear relationships. Therefore, the nonlinear function is better to model actual situations *in vivo*, but it is more difficult to handle and has many more parameters.

According to needs, users can choose the function of corresponding complexity, and select line charts to observe the time-series gene expression data distribution of a sample.

Disease gene selection by significance analysis of microarrays (SAM)

SAM identifies genes with statistically significant changes in expression by ordering a set of gene-specific t-tests. Using permutations, the observed order statistics are compared to those order statistics under the null hypothesis of no differential expression among any genes (Tusher et al., 2001). We used SAM to test our simulation data for each model, and the results are shown in Figure 3. When only one disease gene is selected, the disease genes of each model are all successfully specified by SAM. When two disease genes are selected, the first four cases are found by SAM. One disease gene in the fifth model with low prevalence was not mined.

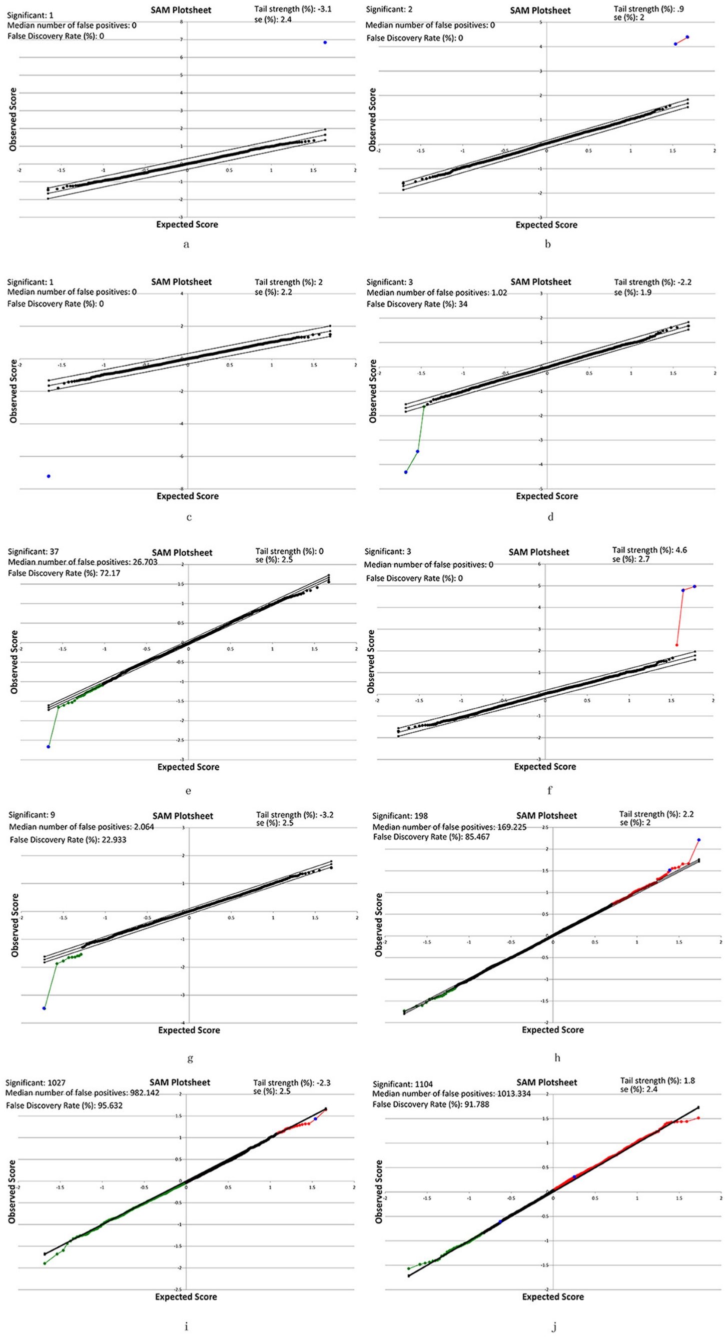


Figure 3. SAM plots of the simulated data. Red and green dots modeled up- and down-regulated genes, respectively. Black dots represent the remaining genes, and blue dots model the disease genes. The first column shows that only one disease gene is selected, and the second column shows that two disease genes are selected. The different rows indicate the different disease models, and the order of the disease model corresponds to Table 3.

DISCUSSION

A simulation method of microarray data is introduced herein to model the occurrence and development of general diseases. According to classical statistics and control theory, five risk models are presented, and prevalence is estimated. One or more model can be embedded into a baseline simulation data set with a case-control label. Additionally, time-series gene expression data can be generated to model a dynamic evolution process of disease. The most significant disease-associated gene is successfully found by “significance analysis of microarrays” (SAM), and individual disease genes with low prevalence are not easy to mine. When the number of disease-related genes increase, the selection of the risk function and the determination of parameters are challenging. In the future, we will continue to develop risk models containing many more disease genes.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Natural Science Foundation of China under Grants #61571341, #61201312, and #91530113; the Research Fund for the Doctoral Program of Higher Education of China (#2013 0203110017); the Fundamental Research Funds for the Central Universities of China (#BDY171416 and #JB140306); and the Natural Science Foundation of Shaanxi Province in China (#2015JM6275).

REFERENCES

Albers CJ, Jansen RC, Kok J, Kuipers OP, et al. (2006). SIMAGE: simulation of DNA-microarray gene expression data. *BMC Bioinformatics* 7: 205. http://dx.doi.org/10.1186/1471-2105-7-205

Bresch D, Colin T, Grenier E, Ribba B, et al. (2010). Computational modeling of solid tumor growth: the avascular stage. *SIAM J. Sci. Comput.* 32: 2321-2344. http://dx.doi.org/10.1137/070708895

Carvajal-Rodríguez A (2008). Simulation of genomes: a review. *Curr. Genomics* 9: 155-159. http://dx.doi.org/10.2174/138920208784340759

Choe SC, Zhao G, Zhao Z, Rosenblatt JD, et al. (2011). Model for *in vivo* progression of tumors based on co-evolving cell population and vasculature. *Sci. Rep.* 1: 31. http://dx.doi.org/10.1038/srep00031

De Smet R and Marchal K (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8: 717-729.

Hoban S, Bertorelle G and Gaggiotti OE (2012). Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* 13: 110-122.

Kuwahara H, Fan M, Wang S and Gao X (2013). A framework for scalable parameter estimation of gene circuit models using structural information. *Bioinformatics* 29: i98-i107. http://dx.doi.org/10.1093/bioinformatics/btt232

Nykter M, Aho T, Ahdesmäki M, Ruusuvuori P, et al. (2006). Simulation of microarray data with realistic characteristics. *BMC Bioinformatics* 7: 349. http://dx.doi.org/10.1186/1471-2105-7-349

Singhal S, Kyvernitis CG, Johnson SW, Kaiser LR, et al. (2003). Microarray data simulator for improved selection of differentially expressed genes. *Cancer Biol. Ther.* 2: 383-391. http://dx.doi.org/10.4161/cbt.2.4.431

Tang W, Wu X, Jiang R and Li Y (2009). Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. *PLoS Genet.* 5: e1000464. http://dx.doi.org/10.1371/journal.pgen.1000464

Tusher VG, Tibshirani R and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98: 5116-5121. http://dx.doi.org/10.1073/pnas.091062498

Vu TT and Vohradsky J (2002). Genexp--a genetic network simulation environment. *Bioinformatics* 18: 1400-1401. http://dx.doi.org/10.1093/bioinformatics/18.10.1400

Yuan X, Miller DJ, Zhang J, Herrington D, et al. (2012). An overview of population genetic data simulation. *J. Comput. Biol.* 19: 42-54. http://dx.doi.org/10.1089/cmb.2010.0188