ExGenet, Integrating Design of Experiments and Response Surface Methodology for Cancer Gene Detection in Gene Regulatory Networks

Cancer Informatics
Volume 23: 1–9
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351241255645
Sage

Mahboube Ayoubi[1], Babak Teimourpour[2] and Alireza Hassanzadeh[3]

[1]Department of Data Science, Tarbiat Modares University (TMU), Tehran, Iran. [2]Department of Information Technology Engineering, School of Systems and Industrial Engineering, Tarbiat Modares University (TMU), Tehran, Iran. [3] Professor and Head of Department of Information Technology Management, Tarbiat Modares University (TMU), Tehran, Iran.

ABSTRACT

OBJECTIVE: Network analysis techniques often require tuning hyperparameters for optimal performance. For instance, the independent cascade model necessitates determining the probability of diffusion. Despite its importance, a consensus on effective parameter adjustment remains elusive.

METHODS: In this study, we propose a novel approach utilizing experimental design methodologies, specifically 2-Factorial Analysis for Screening, and Response Surface Methodology (RSM) for parameter adjustment. We apply this methodology to the task of detecting cancer driver genes in colorectal cancer.

RESULT: Through experimental validation of colorectal cancer data, we demonstrate the effectiveness of our proposed methodology. Compared with existing methods, our approach offers several advantages, including reduced computational overhead, systematic parameter selection grounded in statistical theory, and improved performance in detecting cancer driver genes.

CONCLUSION: This study presents a significant advancement in the field of network analysis by providing a practical and systematic approach to hyperparameter tuning. By optimizing parameter settings, our methodology offers promising implications for critical biomedical applications such as cancer driver gene detection.

KEYWORDS: Independent cascade model, gene regulatory network, experimental design, complex network

RECEIVED: January 22, 2024. **ACCEPTED:** April 26, 2024.

TYPE: Original Research Article

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Babak Teimourpour, Department of Information Technology Engineering, School of Systems and Industrial Engineering, Tarbiat Modares University (TMU), Tehran, 1411713116, Iran. Email: b.teimourpour@modares.ac.ir

Introduction

Social networks have played an important role in spreading information, opinions, ideas, innovations, and rumors around the world.[1] For this reason, several subjects are investigated through the analysis of societal networks, such as diffusion and influence models. One of the most widely used techniques for modeling diffusion processes is a cascade model. The active node v may attempt once to activate one of its adjacent active nodes with a probability of p_v in cascading models beginning with seed nodes and continuing each step t. If it works, the new nodes will be launched in step $(t+1)$ and an identical operation shall be performed for each inactive node to activate them. Active nodes cannot attempt 2 activations of the same node, whether they are successful or not. This process will continue until a new node can no longer be activated.[2-6]

A number of social network analysis works have recently explored the diffusion of information. A widely prevalent basic probabilistic model of information propagation through networks is the independent cascade model (IC model). The IC model assumes that the activation of one node in the network does not influence the activation of another node, which may not always hold true in real-world scenarios. Moreover, the model assumes a fixed probability distribution for the propagation of information or influence through the network. The performance of the IC model can be sensitive to the choice of parameters, such as the propagation probabilities or the seed nodes selected for activation. A set of relevant parameter values should be provided in advance for the IC model. However, knowing the likelihood of diffusion via links for a given network in advance is typically difficult. It is therefore a key research question to identify diffusion probabilities using links from an observed set of data about information dissemination

Saito and colleagues have been studying the parameters of an independent cascade model. They used the EM algorithm to define the problem as a probability function and estimate the probability of the links.[7] A timely asynchronous independent cascade model, where the diffusion probability depends on the timing of transport, was suggested by Guille and Hacid. Using machine learning methods, they were able to infer the diffusion probabilities based on features from Twitter's Social

Network. These studies do not reveal the diffusion probability for each connection and model based on the user's extracted features.[8] Wang et al[9] learned the probability of diffusion in the IC model by assuming that the message is spread between 5 emotions and the diffusion probability differs from emotion to emotion. The issue of learning the diffusion probabilities for IC models is dealt with by Mashayekhi et al. They propose a weighted method for the estimation of diffusion probabilities, which takes into account an information from all previous cascades within the network. Consequently, it is scalable against baselines, resulting in Mashayekhi et al.[10]

Even though there have been studies in the literature on IC model optimization using optimization techniques such as EM algorithms, diffusion probability is still tuned by a trial and error approach. GenIC used an IC model for the detection of cancer genes in colorectal cancer. In this study, diffusion probabilities were chosen by trial and error.[2] The independent cascade model edge probability p_v is usually selected by trial and error; but selecting p_v properly can improve the model to get maximum influence in the network. Experiment design is a statistical method to reach the best value of variables with fewer experiments.[11] The underlying mechanisms of cancer, which may help to identify it more accurately, can be revealed in the regulatory network modules when analyzed quantitatively.[12-14] The identification of cancer genes and their regulation is an important area of research in cancer systems biology.[15-17] In this study, we try to optimize the value of p_v for the independent cascade model. The proposed model will be applied to predict cancer driver genes in the regulatory network for colorectal cancer.

Methods

Gene regulatory network

Gene regulatory network events are of critical importance to various physiological and developmental processes within cells, where the macromolecules, such as genes and ribonucleic acid, are coordinated to create operational responses to a variety of conditions. A disruption in the regulatory relationship between molecules within a cell is one of the causes of cancer. Therefore it is possible to identify the cause of the disturbance through the study of the regulatory network.[18] In order to build the gene regulation network in colorectal cancer (Figure 1), it was necessary to establish interactions and gene expression data (Table 1). We have used the dataset that was also used in Akhavan-Safar et al.[2]

Response surface methods (RSM)

Design of Experiment (DOE) is a method that allows experimenters to organize their experiments and identify the relationships between causes and effects. As it cuts down on the number of tests needed, DOE is widely used in multidisciplinary scientific fields.[19]

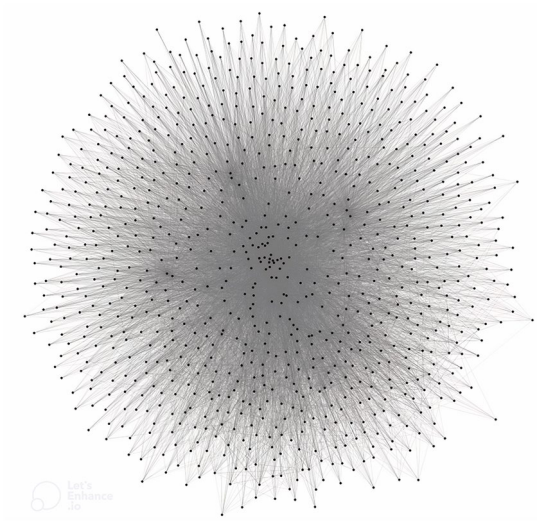

Figure 1. Network of the colorectal cancer gene regulatory network.

Table 1. The characteristics of the data obtained from the RegNetwork.[2]

NUMBER	DESCRIPTION	ELEMENT
21175	All nodes used in the construction of the gene regulatory network	Node
150202	All regulatory interactions used in the construction of the gene regulatory network	Edge
1456	Transcription factors used in the construction of the gene regulation network	TF
19719	target genes used in the construction of the gene regulation network	Gene
149841	The "TF-gene" regulations used in the construction of the gene regulation network	TF-gene
361	The "TF"-"TF gene" self-regulations used in the construction of the gene regulation network	TF-TF

As good ways to optimize process parameters, a number of experimental designs have been selected. RSM designs come in a variety of forms, including factorial design, central composite design (CCD), Box Behnken Design. and D-optimal design.[20]

Analysis of variance (ANOVA) is necessary to verify the model's relevance and fitness since it clarifies whether the developed quadratic model has any real-world importance. It looked into the impact of process parameters and how they interacted. The propagation of error (POE) is taken into account to ensure the robustness of RSM designs. In experiments where uncontrollable components (noise) are assumed

to be zero, POEs—a measurement of the standard deviation of transmitted variability in the output response—are caused by changes in key controllable process variables.[21]

RSM makes the assumption that, within the experimental region, the relationship between the input and response variables is linear and constant. The best results from RSM are obtained from experimental designs that have a continuous response variable and few input variables. For intricate systems with lots of interacting components, it might not be appropriate. The accuracy of the fitted response surface model may be impacted by experimental mistakes or noise in the response variable, as RSM is sensitive to these factors.

The RSM experiment provides a numerical model or equation describing the reaction as an expression of individual factors and levels. A few key pieces of information related to the system under investigation can be identified within this mathematical equation; for example, major factors, factor interactions, and curvatures that indicate whether a response is linear in nature are identified. Another important indicator obtained from a mathematical model based on RSM is the change in direction and magnitude of factor levels to explore an emerging area for improved responses. This process improvement can be achieved using a model based on the steepest ascent. The steepest ascent will allow experimental progress to be made in specific directions with a view to assessing the potential for improving model performance. It is crucial for the experimenter to use all information collected in this model that describes a definite response, rather than relying upon suspicion or speculating about what experimental conditions will be applied in the future.

The starting conditions of an area that proves to have better test performance in the optimization phase of response may not differ much from a more desirable region. An introductory design that is essentially 2 levels of factorial test with repeated focal points may be used in the case of a linear RSM first-order model. However, 2 factors, 2 levels of factorial with center points, as shown in Figure 2, are also applied to a simple 5-point design. If there are only 2 important factors that have an impact on the required response. However, as is usually the case if the improved response region is not close to the original starting conditions, the amount and indication of the direct expressions in the equation below can be adjusted to specify the steepest ascent on the route which leads to 1. If the preliminary RSM test's fitted model is linear, the response area Y shall be improved.

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots b_k x_k \quad (1)$$

Where b-value are parameter estimates that are unaffected by the factors' scaling rule and characterize the scale and direction of the effects, and x_i's represent significant factor effects that are coded. Alternatively, if the desired response calls for minimization rather than maximization, it would be necessary to descend the path by the steepest descent. The units of

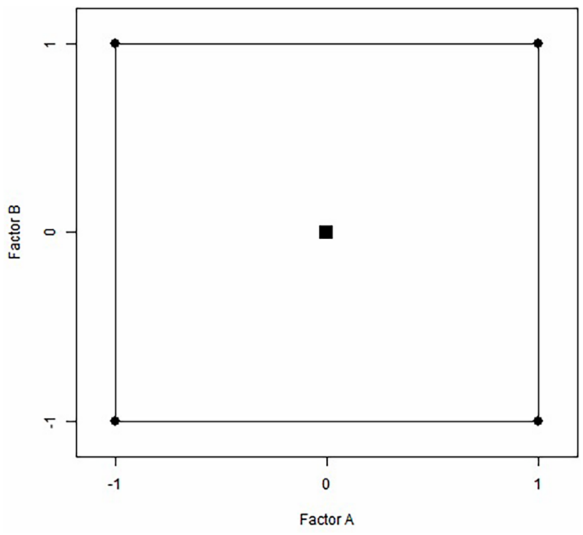

Figure 2. Utilizing the duplicated center and corner points, a simple RSM design. The 2 factors, probability diffusion and iteration, have high, low, and mid-level, respectively, and are represented by +1, –1, and 0, in that order. Square symbol indicates the center points.

Table 2. Levels of the RSM factor (+1/–1) and the zero state at the center.

REGENT	–1	0	1
Probability	0.3	0.4	0.5
Iteration	150	200	250

measurement are removed by using coded variables, facilitating model interpretation. The lowest and highest values of all these factors are set at –1 and 1 respectively to decode the variables, while the midpoint level is coded as a zero value.[22-24]

Two-factor factorial design

A custom 2-level, 2-factorial design with 5 replicated center points has been used to assess the absence of fit, in order to test the surface response and find a first order model. The 5-point design supplies information about possible curvature of the response, which was unable to be retrieved from a 2-level factorial.

Considering the GenIC model, we choose factor levels as shown in Table 2. The response variable is the F-measure for the cancer driver gene (CDG) that the model can detect correctly. Using R and Python, experiments have been designed, statistical analyses have been performed and model predictions have been made.

Path of steepest ascent

The first-order linear model forecast from the RSM experiment, which stated the connection between the response and the factors, was used to create a new set of experimental steps for determining the factor concentration for the experiment in

Table 3. The 2-factor RSM experiment's findings. R software was used to create a customized RSM design that altered probability diffusion and iteration at 3 different levels (−1, 0, +1). Calculation of the target value for the F-measure.

RUN	CODED		UNCODED		F
	PROBABILITY	ITERATION	PROBABILITY	ITERATION	
1	0	0	0.4	200	0.2122
2	1	1	0.5	250	0.2128
3	0	0	0.4	200	0.2118
4	−1	−1	0.3	150	0.2113
5	−1	−1	0.3	150	0.2120
6	1	1	0.5	250	0.2113
7	−1	1	0.3	250	0.2097
8	0	0	0.4	200	0.2121
9	1	−1	0.5	150	0.2142
10	1	−1	0.5	150	0.2152
11	−1	1	0.3	250	0.2041
12	0	0	0.4	200	0.2140
13	0	0	0.4	200	0.2125

the steepest ascent direction. In the fitted first-order model, it was necessary to take stages of sizes commensurate with the values of the |bi| parameter to reach a more optimal position on the steepest ascent route, in order to respond. Direction is according to the sign of b values.

The distance in steps from the RSM experiment's center point was defined by a proportionality constant (ρ). The value of ρ is chosen by the experimenter and is frequently simply set to 1.[25,26] To prevent achieving the value outside of the range of the probability of diffusion (0,1) in just a few steps, in this investigation, it was set to 0.5.

Experimental runs through the path

The route began at the RSM design space's center and served as the starting point for the test's steepest ascent before branching off to explore the surrounding area. To create a set of experimental runs with various diffusion probabilities and iterations of the IC model, an order of steps with equal spacing along the path was chosen. According to the sign of the b-value terms in the equation of the linear model, factors in each step grew or decreased. By calculating the b1 and b2 parameters in the linear model's equation, the amount of rise and decrease of the factors was determined.

Results

Response surface methods (RSM)

Factor RSM design. The 2 factors, diffusion probability and iteration, were varied at 3 levels (coded as −1, 0, and 1) using an RSM model. For each set of conditions, tests were randomly conducted with 3 values of diffusion probability (high, medium, and low; specifically, 0.3, 0.4, and 0.5) and 3 levels of iterations (high, medium, and low; specifically, 150, 200, and 250), resulting in a total of thirteen runs. The response was the F-measure for the cancer driver gene detected in each trial. For the fitting of models and for the steepest ascent, this response was analyzed at all standard levels. The response values were calculated using the 2 standard levels of factors with 2 repetitions for each 1 plus 5 center points that are shown in Table 3.

The model fit assessment for the response was conducted using R software. The overview of the system's results suggested that a 2-factor model should be applied, based on a significant F-statistic. The linear fit has also been deemed to be a simpler and more practical option when determining the steepest ascent. Table 4 display the analysis of variance (ANOVA) of the linear model, which has been proven to be statistically significant. A significant main effect was found for the probability factor and the iteration factor, with statistically significant *P*-value of .0062 and .0110 respectively, so the linear model is an appropriate model here.

Equations (2) and (3) show the predicted response F-measure of the detected cancer driver gene, coded in and without code forms as a function of x1 and x2. The factors may consist of a wide range of measurement units, and the codes make it possible to easily compare coefficients according to their scale These factors have low and high levels denoted by −1 and +1. The coefficients are simply converted into their natural units by the no-code form.

In terms of the coded factors, the final equation:

Table 4. Analysis of variance (ANOVA) for the fitted linear model.

SOURCE	DF	SUM SQ	MEAN SQ	F VALUE	*P*
X_1	1	3.360e-5	3.360e-5	12.39	.0062
X_2	1	2.758e-5	2.758e-5	10.17	.0110
X_1X_2	1	2.090e-6	2.090e-6	0.77	.4030
Residuals	9	2.441e-5	2.710e-6		
F-statistic		11.54			.0025
R-squared		0.6978			
Adjusted R-squared		0.6373			

$$y = 0.2118 + 0.0020x_1 - 0.0019x_2 \quad (2)$$

As regards no coded factors, the final equation:

$$y = 0.2110 + 0.0205\zeta_1 - 0.00004\zeta_2 \quad (3)$$

The IC model probability and iteration numbers are shown in coded form as x1 and x2. The probability needed to be increased in order to maximize response F, as demonstrated by the rising b-value for x1 (0.0020). due to the negative x2 b-value (–0.0019), the number of iterations would be reduced in order to maximize the response. According to the higher b-value for that phrase, the probability had the largest impact on the target.

Calculate the path of the steepest ascent. The steepest ascent path, with the slope b2/b1, represents a track that runs from the center point of the RSM design space. The experiment needs to shift positively by 0.00020 units with each 0.0019 unit moving in a negative direction, resulting in an expected improvement of response for the linear model's coded variables. The proportionality constant, ρ, is usually set to 1 for one of the variables, and commonly, it is the variable with the biggest b-value parameter. Since x1 has a larger parameter estimate in this scenario, the step sizes of other variables are determined in relation to x1, as demonstrated in equation (4).

$$\Delta x_j = \frac{b_j}{b_i / \Delta x_i} \quad j = 1.2 \ldots . k. \, i \neq j \quad (4)$$

If ρ = 1, then x2 (iteration) is altered by b2/b1 units for every unit of change in x1 (probability). In other words, if the step length of probability in coded units is 1, then the associated iteration step size in coded unit is Δx2= 1.03. In this case, however, it was warranted to have a smaller step size than 1 in order to derive the appropriate coordinate values along the path due to the probability's linearity between 0 and 1.

Table 5. A series of test runs has taken place on the steepest ascent. Be aware that for each step, the iteration is dropped by 26 and the diffusion probability is increased by from the starting point (Run 1).

RUN	PROBABILITY	ITERATION	F
14	0.45	174	0.2100
15	0.50	148	0.2079
16	0.55	122	0.2068
17	0.60	96	0.2047
18	0.65	70	0.2041
19	0.70	44	0.2071
20	0.75	18	0.2057

For shorter step lengths the proportionality coefficient was lowered to 0.5, which is as follows:

$$\Delta x_1 = \mathbf{0.5} \quad (5)$$

$$\Delta x_2 = \left(\frac{b_2}{b_1}\right) * \mathbf{0.5 = 0.51} \quad (6)$$

It is appropriate to revert to no-code units for the calculation of the real value that will be used in the highest climb test. Since the step sizes applicable for both x1 and x2 have been known in the RSM study, it is possible to calculate the likelihood of using coating and detector reagent by way of this method:

$$\Delta\zeta_1 = \Delta x_1 * c = 0.5 * 0.1 = 0.05 \quad (7)$$

$$\Delta\zeta_2 = \Delta x_2 * d = 0.51 * 50 \approx 26 \quad (8)$$

These RSM experiments currently operate under the following operating conditions. "c" refers to the step size for diffusion probability from a central point. "d" indicates, at the central point of the current RSM experiment operation, the step size for iteration number.

In other words, during the steepest ascent, with the RSM experiment's origin as the starting point, the probability should be raised by 0.05 for each step, and the iteration should be reduced by 26 as a result.

These calculations, as shown in Table 5, have been applied to design and run a number of experimental runs within the route for the steepest ascent.

Figure 3 plots the F-measure for each step as well as a path leading to the steepest ascent. Despite the variation in the results, all subsequent steps resulted in a decrease in yield, with the maximum value being shown in the first step. Therefore, another first-order model should be used in the general vicinity of that point.

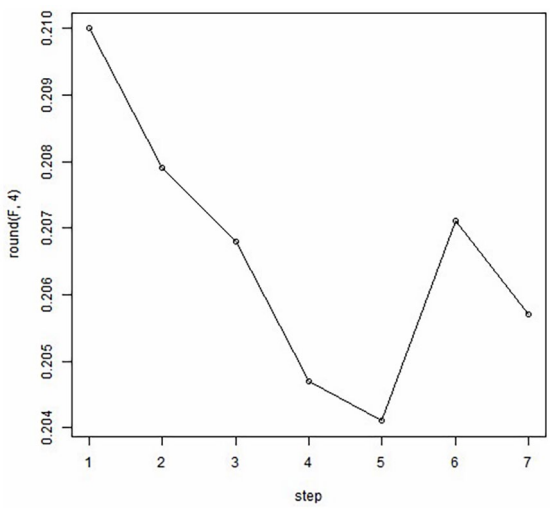

Figure 3. F-measure versus steps during the path of steepest ascent.

A new first-order model is fitted around the point ($\zeta_1 = 0.045$. $\zeta_2 = 174$). The region of exploration for ζ_1 is [0.45,1). Once again, a 2^2 design with 5 center points is used. The analysis of variance shows that the model with a square term is significant (Table 6). The value of the coefficient of determination of model R2 was sufficient (R-Sq = 94.36%), adjusted R^2 value was reasonably in agreement (R-Sq(Adj) = 92.94%). Hence, for the purpose of forecasting, a quadratic model is sufficient.

Diagnostic plots have been obtained and are given in Figure 4, which was used to evaluate how well the model matched the actual response surface. A normal distribution was observed in the residual probability plot, indicating that an acceptable model had been established for use, and there was no pattern when plots of residual versus fitted values were drawn. The Shapiro-Wilk test also shows a *P*-value of .027073, which confirms that residuals are normal.

Figure 5a depicts a contour plot of the response, and Figure 5b display the response surface in 3 planes as a plane that occupies the area between the low and high levels of the 2 factors. As displayed in Figure 5a, we can achieve an F-measure greater than 0.22 around the point ($x_1 = 0.2, x_2 = 0.6$) in the coded value. Computing the global optimum for F-measure gives the value 0.222 in point ($\zeta_1 = 0.47$ $\zeta_2 = 204$).

Colorectal cancer-driven gene detection

Using RSM for tuning IC model parameters has improved it. The colorectal cancer gene regulatory network is the method's input, and the amount of coverage for each gene is the method's output. The number of genes in the gene regulatory network that can be impacted and activated if a gene is active is determined by its coverage in the network. Accordingly, genes that have greater coverage have a higher chance of causing cancer. The study utilized TCGA-COAD, a free TCGA data portal, to identify colorectal cancer genes, with CGC-approved driver genes as the gold standard for evaluation. To evaluate the proposed method, we used common metrics for recall, precision and f-measure performance in binary classification problems. Table 7 display the results of ExGenet for comparison with GenIC and 19 other models, which has a better F-measure as it is maximized by the RSM method. Due to the maximized F-measure, the number of cancer driver genes that the model can detect has improved.

Discussion

Our study employed response surface methods (RSM) to optimize proliferation probability in an independent cascade (IC) model aiming to enhance the detection of cancer driver genes in colorectal cancer through 2-factor RSM implementation plan. Changes in iteration levels were allowed at 3 different levels The results demonstrate that RSMs are both feasible and effective for improving model performance, as evidenced by the analysis of the F-measure as a response variable.

The fit of the linear model was assessed using analysis of variance (ANOVA), which revealed significant statistical results for probability and repeat factors This indicates that the linear model is fitting when the response variable (F-measure) dependent factors are controlled. The observed relationship between proliferation probability, recurrence, and F-measure provides valuable insight into the dynamics of information diffusion in genetic regulatory network.

The identification of a steep ascension path highlights the importance of varying the proliferation probability and repeat level to obtain optimal model performance Our study calculated the steepest ascension path in RSM testing to determine the optimal combination of proliferation probability and repeat for maximize the provided F-measure We evaluated the performance of our proposed ExGenet method for detecting colorectal cancer driver genes and compared it with other existing methods. The results demonstrate that ExGenet identifies a higher number of known driver genes and achieves a better F-measure, indicating its effectiveness in cancer gene discovery. Specifically, ExGenet outperformed the GeneIC method in terms of both the number of known driver genes detected and the F-measure. This validation highlights the importance of optimizing the propagation probability through RSM to improve the performance of the IC model.

The findings of our study hold important implications for cancer genomics research, particularly in the identification of genes causing cancer. Our approach of maximizing the F-measurement through RSM optimization offers a promising approach to improve the accuracy and reliability of cancer gene discovery. Furthermore, the mechanism demonstrated in this study may have broad applications in other cancer types and research areas, paving the way for further advances in cancer genomics Although our study focused on colorectal cancer, future research will investigate the generalizability of our approach to other cancer types and research areas. Additionally, further modifications and extensions of our optimization

Table 6. Analysis of variance (ANOVA) for the fitted quadratic model.

SOURCE	DF	SUM SQ	MEAN SQ	F	*P*-VALUE
A	1	9.1e-6	9.1e-6	3.674	.0697
B	1	4.176e-4	4.176e-4	169.477	<.001
A^2	1	9.52e-5	9.52e-5	38.649	<.001
B^2	1	2.952e-4	2.952e-4	119.784	<.001
AB	1	6.7e-6	6.7e-6	2.729	.1141
Residuals	20	4.93e-5	2.5e-6		
F-statistic	66.86				<0.001
R-squared	0.9436				
Adjusted R-squared	0.9294				

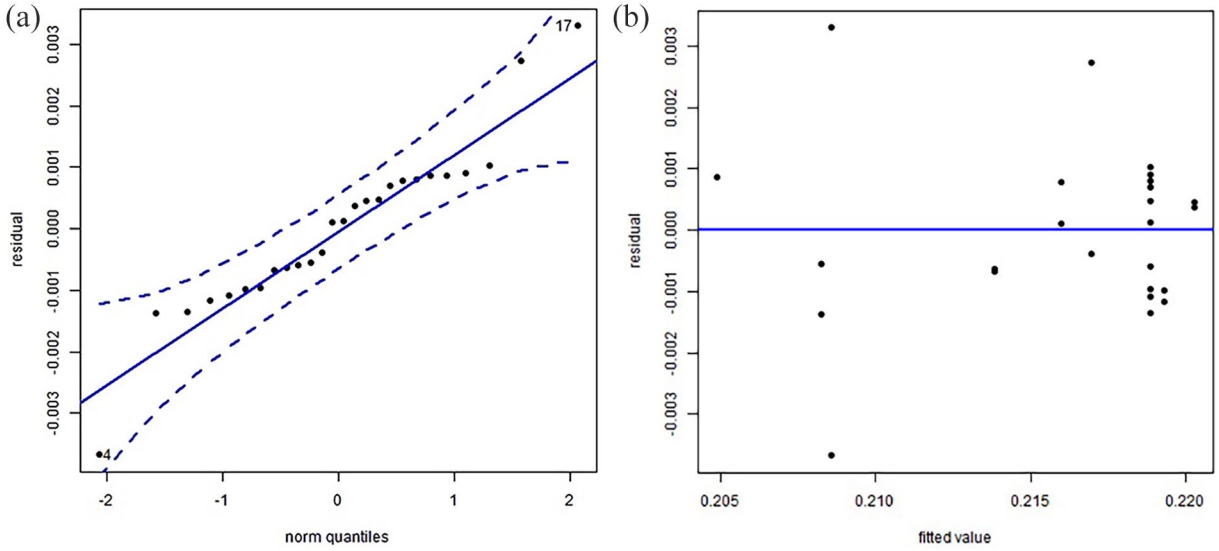

Figure 4. Diagnostic plots displaying (a) the residuals' normal probability plot and (b) the relationship between the residual and the fitted value.

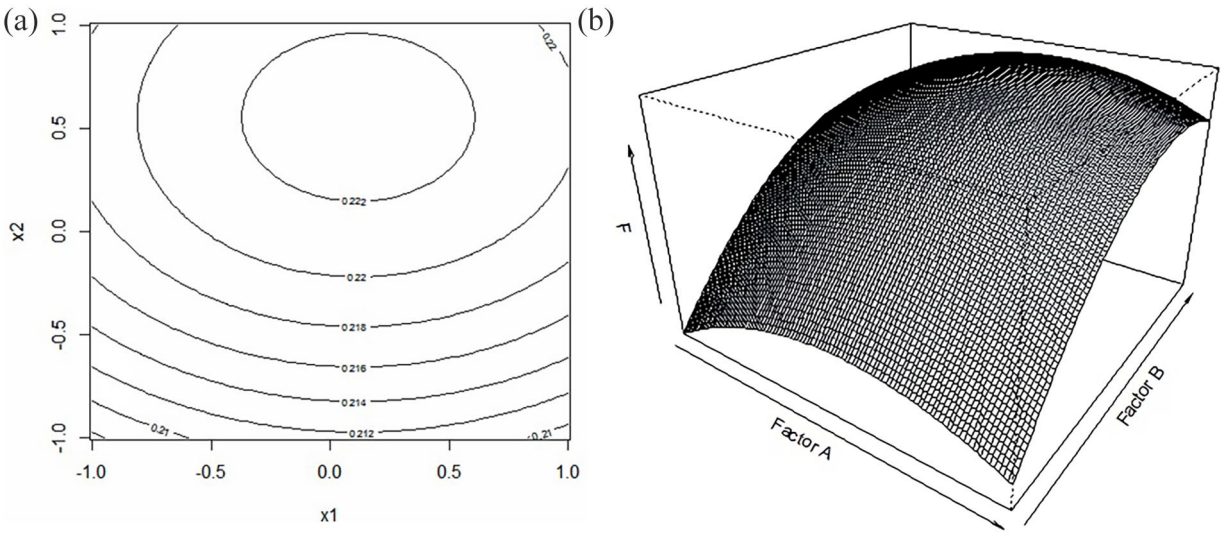

Figure 5. (a) Contour plot of the F response and (b) the response surface plane, in 3-dimensional rendition.

Table 7. Compare ExGenet result with other methods.

METHODS	NUMBER OF DRIVER DETECTED	F-MEASURE	PRECISION	RECALL
NetBox	14	0.043	_	0.024
Simon	46	0.141	0.561	0.08
Dendrix	9	0.024	0.051	0.016
DriverNet	16	0.05	0.254	0.028
MDPFinder	2	0.007	0.333	0
MEMO	0	0	0	0
OncodriveFM	6	0.019	0.115	0.01
ActiveDriver	20	0.059	0.185	0.035
iPAC	286	0.088	0.048	0.5
MutsigCV	8	0.027	0.296	0.014
OncodriveCLUST	5	0.016	0.135	0.009
CoMDP	0	0	0	0
DawnRank	15	0.05	0.5	0.026
e-Driver	18	0.049	0.109	0.031
MSEA	38	0.068	0.07	0.066
DriveML	8	0.027	0.348	0.013
iMaxDriver-N	90	0.194	0.336	0.157
iMaxDrive-W	113	0.214	0.233	0.198
GenHIST	137	0.231	0.221	0.241
GenIC	190	0.218	0.163	0.332
ExGenet	197	0.222	0.164	0.344

method may include considering additional factors or integrating multi-omics data for more detailed analysis.

Limitation

Although valuable insights into colorectal cancer are provided using The Cancer Genome Atlas (TCGA) gene database, there are limitations to potential biases with data collection and its validity, and of patient group representativeness.

Furthermore, although the independent cascade (IC) model is widely used to model diffusion processes, it may oversimplify the complex dynamics of information propagation in gene regulatory networks, leading to errors possible results of predicting diffusion probabilities.

Conclusions

In this study, we propose an experiment design to determine the optimal diffusion probability in the IC model. We have demonstrated that RSM experiments are a feasible and efficient method for quickly exploring an area, in addition to the established design space, by using the steepest ascent approach. This approach takes advantage of a mathematical model that allows the experimenter to map an objective response and improve it.

As we have seen, statistical models like RSM can be utilized to tune parameters and improving the model performance. Additionally, choosing seed nodes in the IC model is another challenge when applying this model. Therefore, employing statistical models to select them could be a subject for further future research.

Acknowledgements

None.

Author Contribution(s)

The contributions of each author to this manuscript are as follows: Mahboube Ayoubi conducted the main aspects of the research, including the conceptualization, methodology, and software (running the code). Additionally, she took the lead in writing the original draft of the manuscript. Babak

Teimourpour, as the first supervisor, played a critical role in the conceptualization of the work, provided overall supervision, and contributed to the writing process by critically reviewing and editing the manuscript. Alireza Hassanzadeh, as the second supervisor, contributed to the supervision of the project and provided valuable input through the review and editing of the manuscript.

Availability of Data and Materials

The datasets analyzed in this study were obtained from The Cancer Genome Atlas (TCGA), a publicly accessible repository for cancer genomics data. Access to TCGA data can be obtained through the official TCGA website (https://www.cancer.gov/tcga) or through the Genomic Data Commons (GDC) Data Portal (https://gdc.cancer.gov/).

Ethics Approval and Consent to Participate

The research presented in this manuscript did not involve human or animal subjects. All data utilized in this study were obtained from publicly available and free-access sources. As such, ethical approval was not required for this research.

Consent for Publication

Given that this study did not involve human or animal subjects and relied on data obtained from free-access websites, informed consent for publication was not applicable in this context. The data utilized in this study were publicly accessible and used in accordance with the terms and conditions of the respective sources.

REFERENCES

1. Banerjee S, Jenamani M, Pratihar DK. A survey on influence maximization in a social network. *Knowl Inf Syst*. 2020;62:3417-3455.
2. Akhavan-Safar M, Teimourpour B, Ayyoubi M. Colorectal cancer driver gene detection in human gene regulatory network using an independent cascade diffusion model. *J Algorithms Comput*. 2022;54:163-185.
3. Angelo GD, Severini L, Velaj Y. Influence maximization in the independent cascade model. In: *Proceedings of the 17th Italian Conference on Theoretical Computer Science (ICTCS16)*, Lecce, Italy, September 7-9, 2016.
4. Berenbrink P, Hahn-Klimroth M, Kaaser D, Krieg L, Rau M. Inference of a rumor's source in the independent cascade model [Internet]. *arXiv*. 2022. Accessed October 9, 2022. http://arxiv.org/abs/2205.12125
5. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. In: *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, June 28-July 1, 2009 [Internet]. Association for Computing Machinery; 2009:199-208.
6. Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: *KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, August 24-27, 2003 [Internet]. Association for Computing Machinery; 2003:137-146.
7. Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model. In: Lovrek I, Howlett RJ, Jain LC eds. *Knowledge-Based Intelligent Information and Engineering Systems*. Lecture Notes in Computer Science. Springer; 2008:67-75.
8. Guille A, Hacid H. A predictive model for the temporal dynamics of information diffusion in online social networks. In: *WWW '12 Companion: Proceedings of the 21st International Conference on World Wide Web*, Lyon, France, April 16-20, 2012 [Internet]. Association for Computing Machinery; 2012:1145–1152.
9. Wang Z, Zhao J, Xu K. Emotion-based Independent Cascade model for information propagation in online social media. In: *13th International Conference on Service Systems and Service Management (ICSSSM)*, Kunming, China, June 24-26, 2016. IEEE; 2016:1-6.
10. Mashayekhi Y, Meybodi MR, Rezvanian A. Weighted estimation of information diffusion probabilities for independent cascade model. In: *4th International Conference on Web Research (ICWR)*, Tehran, Iran, April 25-26, 2018. IEEE; 2018:63-69.
11. Midilli YE, Parsutins S. Optimization of deep learning hyperparameters with experimental design in exchange rate prediction. In: *61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, Riga, Latvia, October 15-16, 2020. IEEE; 2020:1-4.
12. Raza K. Fuzzy logic based approaches for gene regulatory network inference. *Artif Intell Med*. 2019;97:189-203.
13. Wani N, Raza K. IMTF-GRN: integrative matrix tri-factorization for inference of gene regulatory networks. *IEEE Access*. 2019;7:126154-126163.
14. Raza K, Alam M. Recurrent neural network based hybrid model for reconstructing gene regulatory network. *Comput Biol Chem*. 2016;64:322-334.
15. Wani N, Raza K. MKL-GRNI: a parallel multiple kernel learning approach for supervised inference of large-scale gene regulatory networks. *PeerJ Comput Sci*. 2021;7:e363.
16. Raza K, Parveen R. Reconstruction of gene regulatory network of colon cancer using information theoretic approach. In: *Confluence 2013: The Next Generation Information Technology Summit (4th International Conference)*, Noida, September 26-27, 2013. IET; 2013:461-466.
17. Raza K. Reconstruction, topological and gene ontology enrichment analysis of cancerous gene regulatory network modules. *Curr Bioinform*. 2016;11:243-258.
18. Liu ZP, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*. 2015;2015:bav095.
19. Lundstedt T, Seifert E, Abramo L, et al. Experimental design and optimization. *Chemometr Intell Lab Syst*. 1998;42:3-40.
20. Keskin Gündoğdu T, Deniz İ, Çalışkan G, Şahin ES, Azbar N. Experimental design methods for bioengineering applications. *Crit Rev Biotechnol*. 2016;36:368-388.
21. Sadi JA. Private: designing experiments: 3 level full factorial design and variation of processing parameters methods for polymer colors. *Adv Sci Technol Eng Syst J*. 2018;3:109-115.
22. Joyce AP, Leung SS. Use of response surface methods and path of steepest ascent to optimize ligand-binding assay sensitivity. *J Immunol Methods*. 2013;392:12-23.
23. Lujan-Moreno GA, Howard PR, Rojas OG, Montgomery DC. Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. *Expert Syst Appl*. 2018;109:195-205.
24. Montgomery DC. *Design and Analysis of Experiments*. John Wiley & Sons; 2017:752 p.
25. Rodriguez Picon LA, Méndez-González LC, Perez Olguin IJC, García Nava PE. A study of stopping rules in the steepest ascent methodology for the optimization of a simulated process [Internet]. 2022. Accessed February 23, 2023. http://cathi.uacj.mx/handle/20.500.11961/24380
26. Fan SKS, Huang KN. A new search procedure of steepest ascent in response surface exploration. *J Stat Comput Simul*. 2011;81:661-678.