

RESEARCH ARTICLE

# Genetic programming based models in plant tissue culture: An addendum to traditional statistical approach

Meenu R. Mridula<sup>1</sup>\*, Ashalatha S. Nair<sup>1</sup>, K. Satheesh Kumar<sup>2</sup>

**1** Department of Botany, University of Kerala, Kariavattom, Thiruvananthapuram, Kerala, India,

**2** Department of Future Studies, University of Kerala, Kariavattom, Thiruvananthapuram, Kerala, India

✉ These authors contributed equally to this work.

\* [m.r.meenakshi@gmail.com](mailto:m.r.meenakshi@gmail.com)



## Abstract

In this paper, we compared the efficacy of observation based modeling approach using a genetic algorithm with the regular statistical analysis as an alternative methodology in plant research. Preliminary experimental data on *in vitro* rooting was taken for this study with an aim to understand the effect of charcoal and naphthalene acetic acid (NAA) on successful rooting and also to optimize the two variables for maximum result. Observation-based modelling, as well as traditional approach, could identify NAA as a critical factor in rooting of the plantlets under the experimental conditions employed. Symbolic regression analysis using the software deployed here optimised the treatments studied and was successful in identifying the complex non-linear interaction among the variables, with minimalistic preliminary data. The presence of charcoal in the culture medium has a significant impact on root generation by reducing basal callus mass formation. Such an approach is advantageous for establishing *in vitro* culture protocols as these models will have significant potential for saving time and expenditure in plant tissue culture laboratories, and it further reduces the need for specialised background.

## OPEN ACCESS

**Citation:** Mridula MR, Nair AS, Kumar KS (2018) Genetic programming based models in plant tissue culture: An addendum to traditional statistical approach. PLoS Comput Biol 14(2): e1005976. <https://doi.org/10.1371/journal.pcbi.1005976>

**Editor:** Jean-Baptiste Durand, Laboratoire Jean Kuntzmann, FRANCE

**Received:** February 23, 2017

**Accepted:** January 15, 2018

**Published:** February 27, 2018

**Copyright:** © 2018 Mridula et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files. Data has also been uploaded to figshare (DOI: [10.6084/m9.figshare.5844765](https://doi.org/10.6084/m9.figshare.5844765)).

**Funding:** The first author (MRM) would like to acknowledge the financial assistance from Council of Scientific and Industrial Research, New Delhi through Junior Research Fellowship scheme (No. 09/102/RENEWAL/2014-EMR-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author summary

Trials to find out the best combination of factors that contribute to the desired response takes up the chunk of time and resources in any plant tissue culture laboratory. The output of such experiments is analysed statistically to come to a conclusion. However, without prior statistical modifications, the results could be misleading. Recent reports from several labs point out the use of artificial neural networks to circumvent this. We have chosen to use a computational process that can predict the best combination of factors for the desired response after randomly testing the higher and lower limit of the factors with experiments. The magnitude of the desired response can be presumed at any concentration within this range using the models generated by symbolic regression. The procedure provides both optimum model function as well as the optimum variable values in the model. The variable sensitivity and percentage response add depth to the information thus obtained. The study indicated that these models would have significant potential for

**Competing interests:** The authors have declared that no competing interests exist.

saving time and expenditure in plant tissue culture laboratories for the commercial establishment of *in vitro* protocols.

## Introduction

Relatively more straightforward and efficient empirical modeling techniques based on input-output models are gaining popularity to conventional statistical methods across various disciplines [1]. This surge is due to its relative ease of use and understanding. Genetic programming (GP) is an approach which uses the concept of biological evolution to handle a problem with many fluctuating variables. Computational optimisation techniques have recently debuted in plant tissue culture research as studied in neural networks models [2]. Symbolic regression was one of the earliest applications of GP and continues to be widely considered [3]. A broad array of scientific fields like Biology, Chemistry, Environmental Science, Neurology and Psychology reports the use of symbolic regression [4–9]. However, plant tissue culture data has not yet been analysed using symbolic regression. The data generated from plant tissue culture experiments includes continuous, count, binomial or multinomial and predominantly the information is validated using analysis of variance method (ANOVA) [2,10]. ANOVA is adequate for normally distributed continuous data; but without prior manipulation, it is erroneous to analyze count, binomial or multinomial data [11]. Neuro-fuzzy logic is the standard practice by which computational modeling is achieved in plant tissue culture [2,12]. In this context, genetic algorithm based symbolic regression remains unevaluated. Unlike conventional regression analysis which optimises parameters for a pre-defined model, symbolic regression avoids imposing any *a priori* assumptions. In generalised linear model (GLM) regression, the dependent variable is represented as linear combination of the given set of basic functions and optimise the coefficients to fit the data. However, symbolic regression searches for both a set of basic functions and coefficients. The added value of symbolic regression, compared to GLM, lies in its ability to quickly and accurately find an optimal set of basic functions [13,14]. The algorithm infers the model from the data by combining variables and mathematical operators and generates an empirical formula which is a mathematical equation that predicts observed results derived from conducted experiments. GP combines previous equations and forms new ones. Thus it produces models with interpretable structure, relating to input and output variables from a data set without pre-processing and identifying critical parameters and hence shed insight into the underlying processes involved in a given system [15]. Symbolic regression can recognise and model complex non-linear relationships between the inputs and outputs of biological processes even in the presence of disturbances and potential for parallel processing. The preliminary data generated from experiments during rooting of *in vitro* regenerated plantlets in *Wrightia tinctoria* was employed to study the utility of symbolic regression to analyze plant tissue culture data. The effect of two variables - NAA and charcoal on root proliferation was considered. The datasets were subjected to usual statistical analysis as well as observation based modeling via symbolic regression. Moreover, we aimed to optimise the process by examining the influencing factors. We propound the use of symbolic regression-based model prediction as an addendum to data analysis method for plant tissue culture experiments.

## Materials and methods

### Culture conditions

The genetic variability was kept minimum by using a single field grown ortet, thus minimising statistical errors [16]. Nodal regions derived from the fresh flushes of growth from the ortet, two

weeks after lopping one major branch served as the explants [17]. The nodal explants were conditioned over a period of 4 months (subculture/four weeks) on MS medium (1962) [18], pH 5.8 and 2  $\mu\text{M}$  each of BAP and NAA for shoot multiplication. For rooting experiments individual shoots were transferred on MS medium containing 2  $\mu\text{M}$  BAP with NAA (2, 4 and 6  $\mu\text{M}$ , respectively) and charcoal (0.01, 0.03, 0.05, 0.07, 0.09 and 0.11%, respectively) in 250 ml culture flasks in 50 ml of sterilized medium (pH 5.8). The cultures were maintained at  $25\pm 2^\circ\text{C}$  in a culture room with  $40\ \mu\text{molm}^{-2}\ \text{s}^{-1}$  irradiances and a photoperiod of 8 hrs with  $55\pm 5\%$  of relative humidity.

### Experimental design

The plant tissue culture database, containing 21 conditions, followed a factorial design for two variables- concentration of NAA (2, 4 and 6  $\mu\text{M}$ ) and charcoal (0,0.01, 0.03,0.05,0.07,0.09 and 0.11%) in the medium. Each treatment consisted of 5-7 explants in a culture flask with three replicates. The subculture was done at the end of four weeks and five parameters were recorded to analyze the effects of the variables on rooting such as basal callus diameter (mm) (BC), the percentage of shoots rooted (R), the length of the longest root (cm) (RL), the number of roots (NR) and the number of lateral roots (NLR) (S1).

### Statistical analysis

All experiments were conducted using Randomised Block Design (RBD). Continuous data were analysed using multiple linear regression in R and posthoc comparisons of pairs performed by Tukey's test ( $p>0.05$ ). Count data were analysed using Poisson regression model. Pearson's Chi-squared test for count data was employed to access statistical significance of the variables.

### Symbolic regression

Each of the observed parameters is modeled as a function of NAA and charcoal concentrations using symbolic regression and GLM for comparison. To obtain a global optimum, we have also modelled the combination (R+RL+NR+NRL-BC) by taking rooting factors together after normalisation by employing both GLM and symbolic regression. The optimum model for each case was generated by genetic programming based symbolic regression using the software package Eureqa (Version 0.98 beta) with 50% of the data randomly selected as training data, and 3-fold cross-validated with randomly selected 25% of the remaining data [19–21]. Corresponding to each symbolic regression model of the data partition, we have also obtained generalised linear model by including  $x$ ,  $y$ ,  $xy$ ,  $1/x$ ,  $1/y$ ,  $\sin(x)$ ,  $\cos(x)$ ,  $\sin(y)$ ,  $\cos(y)$ ,  $xy \sin(x)$ ,  $xy \cos(x)$ ,  $xy \sin(y)$ ,  $xy \cos(y)$  into the set of basic functions and cross-validated similarly. The remaining 25% the data was used for testing and reporting error [19–21]. The Target expression used to generate the regression model was the minimal equation  $z = f(x, y)$  where 'x' corresponds to NAA concentration and 'y' corresponds to charcoal concentrations, and 'z' represents each of the five observed parameters and their combination. The models were based on the primary and trigonometric building blocks, with the  $R^2$  goodness of fit as the error metric [22,23]. Root Mean Squared Error (RMSE) was calculated for the test data sets. Sensitivity represents the relative impact of the variable on the parameter studied within this model and was calculated by the local method using the partial derivatives [24]. Given a model equation of the form  $z = f(x, y)$ , the influence metrics of x on z was;

$$\text{Sensitivity} = \left| \frac{\partial z}{\partial x} \right| \cdot \frac{\sigma(x)}{\sigma(z)}, \text{ evaluated at all input data points;}$$

**Table 1. Mean of basal callus diameter after four weeks of culture under the conditions studied.** Means followed by similar letters; a, b and c (within each column) and A and B (in each row) are not significantly different by Tukey’s multirange test ( $p>0.05$ ).

Basal callus Diameter (mm)	Charcoal	NAA		
		2	4	6
	0	10.06 c A	11.40 d B	11.33 c B
0.01	2.25 b AB	2.40 c B	2. b A	
0.03	1.61 a A	2.38 c B	1.43 a A	
0.05	1.57 a A	1.54 b A	1.32 a A	
0.07	1.32 a A	1.36 ab A	1.22 a A	
0.09	1.34 a A	1.12 ab A	1.27 a A	
0.1	1.44 a A	1.28 a A	1.24 a A	

<https://doi.org/10.1371/journal.pcbi.1005976.t001>

The percentage positive was calculated as percentage of data points where  $\frac{\sigma(x)}{\sigma(z)} > 0$  and percentage negative was calculated as percentage of data points where  $\frac{\sigma(x)}{\sigma(z)} < 0$ ; where  $\frac{\partial z}{\partial x}$  was the partial derivative of z with respect to x,  $\sigma(x)$  was the standard deviation of x in the input data,  $\sigma(z)$  was the standard deviation of z,  $|x|$  denoted the absolute value of x, and  $\bar{x}$  denoted the mean of x [25]. The ‘fmin’ function in MATLAB (R2012b) was used to obtain the maximum value of each of these functions.

## Results and discussion

The average values obtained for the five growth parameters observed during the study were given as the basal callus diameter (Table 1), the percentage of shoots rooted (Table 2), the length of longest roots (Table 3) and the number of roots and the number of lateral roots (Table 4). The miniscule alphabets within a column indicated the significant influence of charcoal and majuscule alphabets in the row represented the significant interaction of NAA. The shoots inoculated on MS medium with 0% charcoal (control) showed maximum basal callus formation (Fig 1). The shoots inoculated on MS medium supplemented with 4µM NAA and 0.07% charcoal showed the maximum percentage of rooting (Fig 2).

Multiple linear regression demonstrated a significant effect of NAA and its interaction with charcoal on basal callus ( $p>0.001$ ), the percentage of shoots rooted ( $p>0.05$ ) and root length ( $p>0.01$ ) (Tables 1–3). The individualistic effect of NAA for the number of roots and lateral roots were found to be significant at  $p>0.05$  and  $p>0.001$  respectively (Table 4). The interaction of NAA and charcoal was not significant for the same parameters studied. Mathematical

**Table 2. Mean of the percentage of shoots rooted after four weeks of culture under the conditions studied.** Means followed by similar letters; a, b and c (within each column) and A and C (in each row) are not significantly different by Tukey’s multirange test ( $p>0.05$ ).

Percentage of shoots rooted	Charcoal	NAA		
		2	4	6
	0	0 a A	0 a A	0 a A
0.01	0 a A	0 a A	0 a A	
0.03	0 a A	10.76 a B	12.96 b B	
0.05	43.9 c B	47.6 c B	12.8 b A	
0.07	39.93 c B	50 c C	28 c A	
0.09	27.43 b B	31.7 b B	9.4 ab A	
0.1	28.3 b B	23.23 b B	0 a A	

<https://doi.org/10.1371/journal.pcbi.1005976.t002>



**Table 3. Mean of the length of longest roots after four weeks of culture under the conditions studied.** Means followed by similar letters; a, b and c (within each column) and A and B (in each row) are not significantly different by Tukey's multirange test ( $p > 0.05$ ).

	Charcoal	NAA		
		2	4	6
	Length of longest roots (mm)	0	0 a A	0 a A
	0.01	0 a A	0 a A	0 a A
	0.03	0 a A	7.88 b B	1.72 a A
	0.05	26.4 c C	17.94 c B	6.9 ab A
	0.07	41.78 d B	41.67 d B	14.4 b A
	0.09	18.17 b B	10.1 b A	6.8 a A
	0.1	7.38 a B	3.13 ab AB	0 a A

<https://doi.org/10.1371/journal.pcbi.1005976.t003>

functions were successfully developed using symbolic regression to understand the correlation between the two variables for each of the parameters considered and is contrasted with those obtained by traditional regression models (Table 5). To analyze the effect of each of the variables on the parameter studied; variable sensitivity measures were calculated along with its percentage impact. Its sensitivity denoted the relative impact within this model that a variable has on the target variable. The individualistic effect of the two input variables on the output parameter was pointed out as percentage positive or negative of that input variable (Table 6). For the parameter basal calli diameter, the percentage positive value for variable 'y' was zero. In other words, there was zero percent chance of basal calli mass increasing with increasing concentration of charcoal; or that basal calli mass decrease with increasing concentration of charcoal (Fig 3). The model predicted that increase in charcoal concentration had a consequent increase in root length and root number in 50% of all the trials while the same promoted rooting percentage and lateral root number in 75% of the trials. Root number and root length decreased with increasing concentration of NAA in 100% of the trials. Rooting percentage and lateral root numbers increased with increasing NAA concentration in 50% of all the trials.

**Table 4. Mean of the number of roots and number of lateral roots after four weeks of culture under the conditions studied.** The row and the column variables are statistically significantly associated at  $p > 0.05$  (Number of roots) and  $p > 0.001$  (Number of lateral roots), by Pearson's Chi-squared Test for Count Data.

	Charcoal	NAA		
		2	4	6
	Number of Roots	0	0	0
	0.01	0	0	0
	0.03	0	0.4	0.6
	0.05	5	2.4	1.2
	0.07	1.5	1.4	0.8
	0.09	1	1.13	0.8
	0.1	0.8	0.6	0
Number of lateral Roots	0	0	0	0
	0.01	0	0	0
	0.03	0	0	0
	0.05	0.4	0.43	0
	0.07	5	0.6	1.1
	0.09	0.6	2.7	2.2
	0.1	0	1.96	0

<https://doi.org/10.1371/journal.pcbi.1005976.t004>



**Fig 1.** *In vitro* plants of *W. tinctoria*- Shoot inoculated in the absence of charcoal showing profuse basal callusing (left); against shoot inoculated in the presence of charcoal showing reduced/no basal callus mass (right).

<https://doi.org/10.1371/journal.pcbi.1005976.g001>

The function obtained and the 3D plots thus generated could be used to predict the combinations of input variables giving optimum results. The best response for rooting percentage



**Fig 2. Maximum rooting observed in the shoots inoculated on MS medium supplemented with 4  $\mu$ M NAA and 0.07% charcoal.**

<https://doi.org/10.1371/journal.pcbi.1005976.g002>

was predicted at 3.7  $\mu$ M NAA and 0.08% charcoal (Fig 4). The root length showed a non-linear pattern, and the highest value for its function was estimated with 2.8  $\mu$ M NAA and 0.05% charcoal (Fig 5). The maximum root number was determined for 1.7  $\mu$ M NAA and 0.06% charcoal (Fig 6). The maximum value of the function generated for lateral root number was with 6.3  $\mu$ M NAA and 0.08% charcoal (Fig 7). The global optimum modelled upon the combination (R+RL+NR+NRL-BC) indicated the results as 2.44  $\mu$ M NAA and 0.03% charcoal (Fig 8).

Table 5. Functions and error matrix obtained through symbolic regression and multiple linear regression.

Parameter	Symbolic regression			Linear regression		
	Function	R <sup>2</sup>	RMSE	Function	R <sup>2</sup>	RMSE
Basal callus diameter (BC)	$1.17 + 0.12/(0.01 + y)$	0.98	0.09	$739.5 x - 51540 y + 73390 \sin(y) - 0.56 \sin(x) - 29450 \cos(y) - 1190 xy + 1673 x \sin(y) - 14130 y \sin(y) + 739.3 x \cos(y) - 21970 y \cos(y) + 0.7174 y \sin(x) - 350.2 xy \sin(y) - 487.8 xy \cos(y) + 294660$	0.98	0.56
Percentage of shoots rooted (R)	$4.43 x + 4738572.57 y^4 + 88353.22 y^2 - 18.33 - 1270650.34 y^3 - 21.85 y x^2$	0.73	0.94	$70990 x + 619700 y - 904900 \sin(y) + 0.294 \sin(x) + 278000 \cos(y) - 157200 xy + 229300 x \sin(y) + 137300 y \sin(y) - 70990 x \cos(y) + 285300 y \cos(y) - 27.01 y \sin(x) - 34940 x y \sin(y) - 72090 x y \cos(y) - 278000$	0.60	1.11
Length of longest roots (mm) (RL)	$2.28 - 22.58 \cos(40.39 y)/(1.25 x + x \cos(5.06 y) + \sin(1.84 - x))$	0.75	1.28	$46840 x + 530000 y - 768300 \sin(y) - 1.7 \sin(x) + 263300 \cos(y) - 98130 xy + 142600 x \sin(y) + 129200 y \sin(y) + 129200 y \sin(y) - 46840 + 238500 y \cos(y) - 4.134 y \sin(x) - 22990 x y \sin(y) - 44490 x y \cos(y) - 263300$	0.85	7.85
Number of roots (NR)	$14.83 y^2 + 14.78 y^2 \sin(10.50 + 6375y) - 4.80 y - 1.92 x y - 19.2 y \sin(5.5 + 88.86y) - 1.95 x y \sin(6.3 - 183.03y)$	0.82	0.76	$5984 x + 56560 y - 83160 \sin(y) + 0.09042 \sin(x) + 22980 \cos(y) - 14110 xy + 20670 x \sin(y) + 11420 y \sin(y) - 5984 x \cos(y) + 26590 y \cos(y) - 0.3519 y \sin(x) - 2959 x y \sin(y) - 6560 x y \cos(y) - 22980$	0.63	1.03
Number of lateral roots (NRL)	$0.04 + 2208.78 x y^3 - 3559.64 y^4 - 0.39 y x^2 - 17758.21 x y^4$	0.96	0.02	$7025 x + 56710 y - 82060 \sin(y) + 0.467 \sin(x) + 28560 \cos(y) - 14700 x y + 21340 x \sin(y) + 13970 y \sin(y) - 7025 x \cos(y) + 25370 y \cos(y) - 2.28 y \sin(x) - 3443 xy \sin(y) - 6652 xy \cos(y) - 28560$	0.68	1.40
Normalised (R+RL+NR+NRL-BC)	$1.24 \sin(3.24 + 3.42 x^3 y^2 \sin(y) - 3.77 y^2 \cos(x))$	0.80	0.92	$8604 x + 15570 y - 19520 xy - 2.067 \sin(x) - 22750 \sin(y) + 6988 \cos(y) + 3453 y \sin(x) - 8606 x \cos(y) - 4249 xy \sin(y) - 9014 xy \cos(y) - 6988$	0.78	1.99

The variable x represents NAA concentration and y represents charcoal concentration

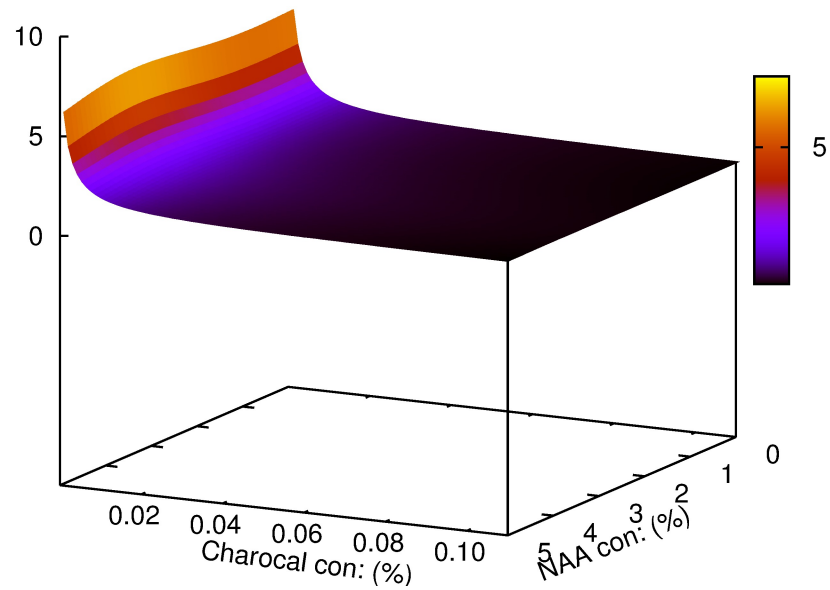
<https://doi.org/10.1371/journal.pcbi.1005976.t005>

The conclusion obtained by traditional statistics suggested that charcoal had a positive and stimulatory effect in rooting of shoots by reducing basal callus (Table 1). Percentage of shoots rooted and root length showed a significant impact with the combination of NAA and charcoal (Tables 2 and 3). In the present study, NAA has a significant effect on rooting as shown by the number of roots and lateral roots (Table 4). Similar results were reported in *Acacia leucophloea* and *Cinnamomum verum* [26, 27]. With traditional statistics, we were not able to estimate the combination/s of both variables in producing the best results or able to identify the relative impact of a particular variable on the output parameter. Modeling of plant tissue culture data is practised using regression analysis where first an initial function is approximated and the data fitted to that function to obtain the optimum parameters [11, 28, 29]. In this procedure even when one gets the optimum parameter values, the model prediction was limited by the probable wrong selection of the model function. In contrast, symbolic regression

Table 6. Variable sensitivity values of the parameters studied through symbolic regression.

Parameters	Sensitivity		Percentage positive		Percentage negative	
	x	y	x	y	x	y
Basal calli diameter (mm)	-	1.65	-	0	-	100
Percentage of shoots rooted	0.7	1.78	50	75	50	25
Length of longest roots (mm)	0.3	1.4	0	50	100	50
Number of roots	0.21	1.26	0	50	100	50
Number of lateral roots	0.18	1.94	50	75	50	25
Normalised	0.13	1.64	50	75	50	25

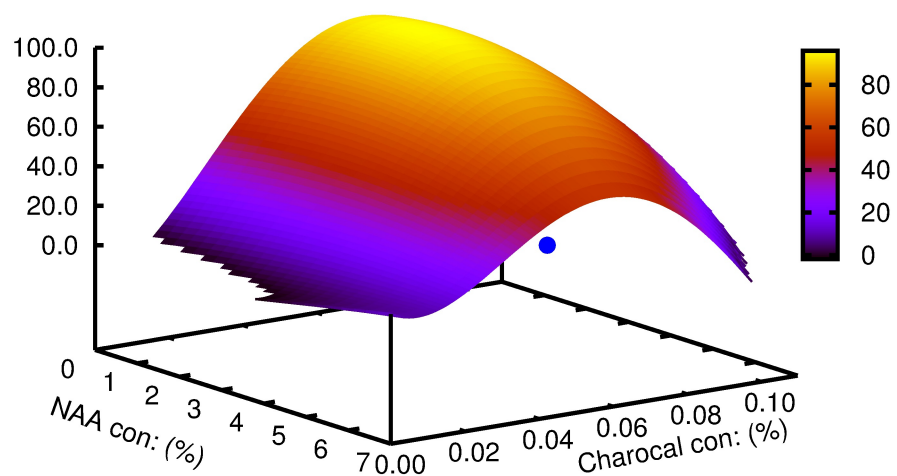
<https://doi.org/10.1371/journal.pcbi.1005976.t006>



**Fig 3.** The plot generated for parameters predicted by symbolic regression model for *W. tinctoria* plantlets as a function of NAA and charcoal concentration showing basal calli diameter (mm). The cyan coloured dot indicates the optimum concentration predicted by the model function.

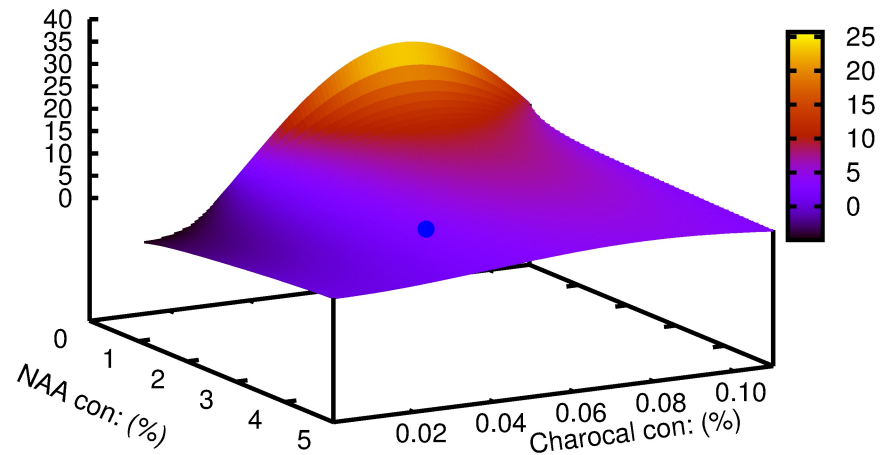
<https://doi.org/10.1371/journal.pcbi.1005976.g003>

procedures work simultaneously on model specification problem and the problem of fitting coefficients [30]. Thus it provides both optimum model function as well as the optimum variable values in the model. The simple relations derived from GP were more accessible to analyze the relationships between the input and output variables [31]. Observation-based predictive models using GP identified that the individualistic effect of charcoal was significant in all the output parameters. A previous investigation suggested basal callus mass formation as one of the primary constraints in the culture of this tree species [32]. In the present study, charcoal has a positive and stimulatory effect in rooting by reducing basal callus formation in shoots. For each of the functions, generated values can be obtained by increasing /decreasing the



**Fig 4.** The plot generated for parameters predicted by symbolic regression model for *W. tinctoria* plantlets as a function of NAA and charcoal concentration showing rooting percentage. The cyan coloured dot indicates the optimum concentration predicted by the model function.

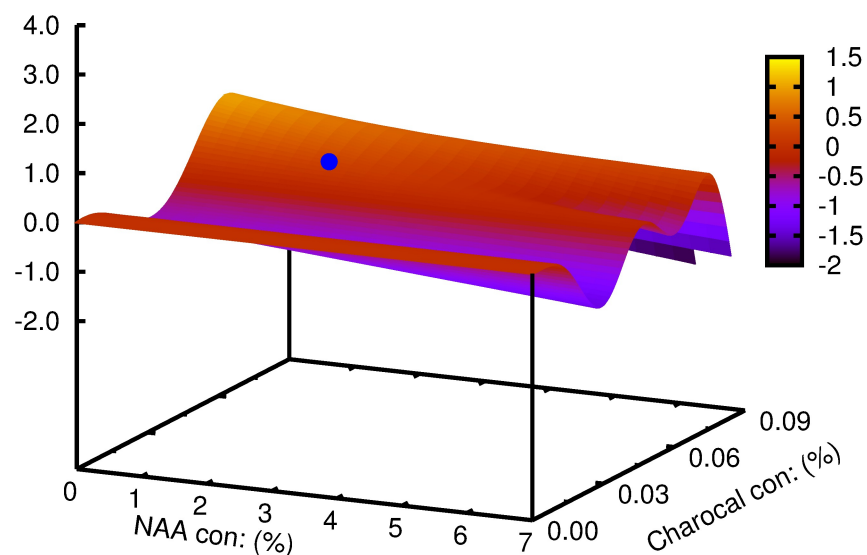
<https://doi.org/10.1371/journal.pcbi.1005976.g004>



**Fig 5.** The plot generated for parameters predicted by symbolic regression model for *W. tinctoria* plantlets as a function of NAA and charcoal concentration showing the length of the longest roots (mm). The cyan coloured dot indicates the optimum concentration predicted by the model function.

<https://doi.org/10.1371/journal.pcbi.1005976.g005>

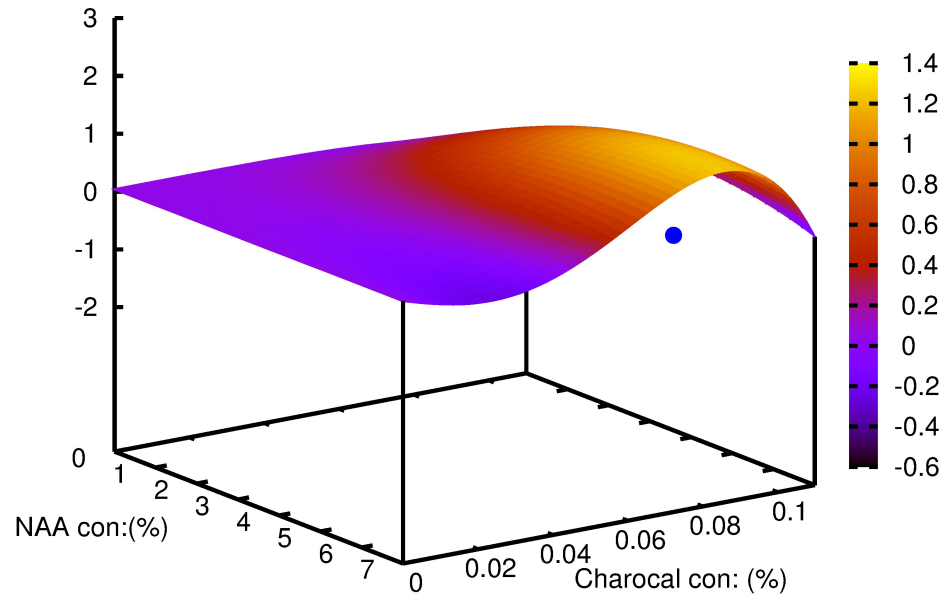
variables by a unit. After randomly testing the higher and lower limit of the additives with experiments, the magnitude of the observed parameters can be presumed at any concentration of the additives within this range using the models generated. It can be extended to analyze synergistic interactions between two parameters by testing whether increasing both variables by a unit, gives a higher or a lower value than the sum of the values obtained by increasing each individually by a unit. The basic requirement for any empirical model includes interpretability, robustness and reliability [33]. Symbolic regression gave comparably lesser RMSE values in comparison to multiple linear regression, thus adding validity to its use. In plant tissue culture obtaining an optimum model is crucial when one needs to find the optimum experimental parameters for large-scale production. The procedure adopted in the work can also be



**Fig 6.** The plot generated for parameters predicted by symbolic regression model for *W. tinctoria* plantlets as a function of NAA and charcoal concentration showing the number of roots. The cyan coloured dot indicates the optimum concentration predicted by the model function.

<https://doi.org/10.1371/journal.pcbi.1005976.g006>

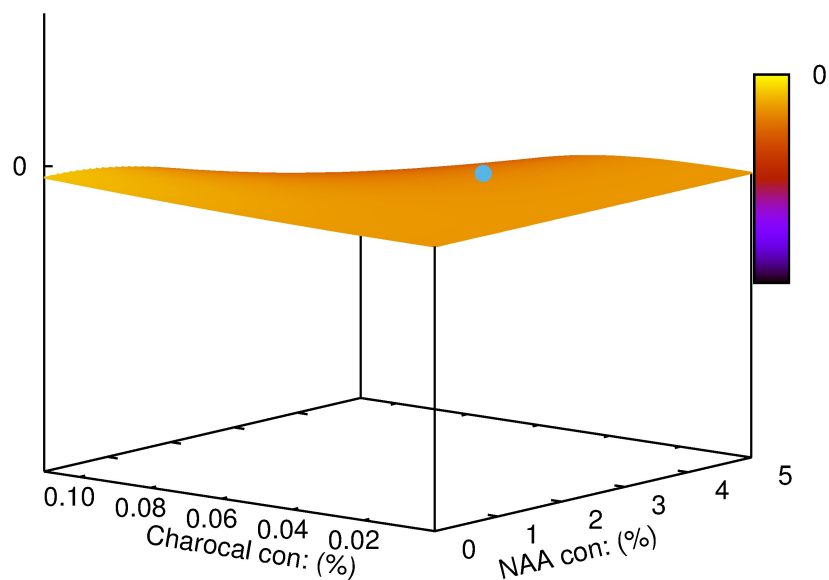




**Fig 7. The plot generated for parameters predicted by symbolic regression model for *W. tinctoria* plantlets as a function of NAA and charcoal concentration showing the number of lateral roots. The cyan coloured dot indicates the optimum concentration predicted by the model function.**

<https://doi.org/10.1371/journal.pcbi.1005976.g007>

extended to similar experiments as it is general and computationally efficient. The analysis predicted the optimum concentration of medium for micropropagation of the selected tree species from the model plots derived from the preliminary experimental data. The study indicated that these models would have significant potential for saving time and expenditure in plant tissue culture laboratories for the commercial establishment of *in vitro* protocols in tree species.



**Fig 8. The plot generated for the combination of all parameters studied, by symbolic regression model for *W. tinctoria* plantlets as a function of NAA and charcoal concentration. The cyan coloured dot indicates the optimum concentration predicted by the model function.**

<https://doi.org/10.1371/journal.pcbi.1005976.g008>

## Supporting information

**S1 Data. The minimal data set used for the analysis.** Legends NAA, CH, BC, R, RL, NR and NRL represent the concentration of NAA, concentration of charcoal, basal callus diameter, percentage of shoots rooted, length of longest roots, number of roots and number of lateral roots, respectively.

(CSV)

## Author Contributions

**Conceptualization:** Meenu R. Mridula, K. Satheesh Kumar.

**Data curation:** Meenu R. Mridula.

**Formal analysis:** Meenu R. Mridula, K. Satheesh Kumar.

**Funding acquisition:** Ashalatha S. Nair.

**Investigation:** Meenu R. Mridula, Ashalatha S. Nair.

**Methodology:** K. Satheesh Kumar.

**Project administration:** Ashalatha S. Nair.

**Resources:** Ashalatha S. Nair, K. Satheesh Kumar.

**Software:** Meenu R. Mridula, K. Satheesh Kumar.

**Supervision:** Ashalatha S. Nair, K. Satheesh Kumar.

**Validation:** K. Satheesh Kumar.

**Visualization:** K. Satheesh Kumar.

**Writing – original draft:** Meenu R. Mridula.

**Writing – review & editing:** Ashalatha S. Nair.

## References

1. Vladislavleva EJ, Smits GF, Den Hertog D. Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *IEEE Transactions on Evolutionary Computation*. 2009 Apr; 13(2):333–49.
2. Gago J, Martínez-Núñez L, Landín M, Gallego PP. Artificial neural networks as an alternative to the traditional statistical methodology in plant research. *Journal of plant physiology*. 2010 Jan 1; 167(1):23–7. <https://doi.org/10.1016/j.jplph.2009.07.007> PMID: 19716625
3. Angeline PJ. Genetic programming: On the programming of computers by means of natural selection: Koza John R., A Bradford Book, MIT Press, Cambridge MA, 1992,819 pp.
4. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. *Nature methods*. 2012 Jun; 9(6):621–625. <https://doi.org/10.1038/nmeth.1975> PMID: 22504588
5. Andrianasolo FN, Casadebaig P, Maza E, Champolivier L, Maury P, Debaeke P. Prediction of sunflower grain oil concentration as a function of variety, crop management and environment using statistical models. *European Journal of Agronomy*. 2014 Mar 1; 54:84–96.
6. Mehrkesh A, Karunanithi AT. New quantum chemistry-based descriptors for better prediction of melting point and viscosity of ionic liquids. *Fluid Phase Equilibria*. 2016 Nov 15; 427:498–503.
7. Xu J, Wang J, Wei Q, Wang Y. Symbolic regression equations for calculating daily reference evapotranspiration with the same input to Hargreaves-Samani in arid China. *Water resources management*. 2016 Apr 1; 30(6):2055–73.
8. Mahouti P, Güneş F, Belen MA, Demirel S. Symbolic Regression for Derivation of an Accurate Analytical Formulation using “Big Data” An Application Example. *Applied Computational Electromagnetics Society Journal*. 2017 May 1; 32(5):372–80.

9. Paul T, Stanley K, Osgood N, Bell S, Muhajarine N. Scaling behaviour of human mobility distributions. In International Conference on Geographic Information Science 2016 Sep 27 (pp. 145–159). Springer, Cham.
10. Bewick V, Cheek L, Ball J. Statistics review 9: one-way analysis of variance. *Critical care*. 2004 Apr; 8(2):130. <https://doi.org/10.1186/cc2836> PMID: 15025774
11. Compton ME. Statistical methods suitable for the analysis of plant tissue culture data. *Plant Cell, Tissue and Organ Culture*. 1994 Jun 1; 37(3):217–42.
12. Gago J, Martínez-Núñez L, Landín M, Flexas J, Gallego PP. Modeling the effects of light and sucrose on *in vitro* propagated plants: a multiscale system analysis using artificial intelligence technology. *PLOS One*. 2014 Jan 20; 9(1):e85989. <https://doi.org/10.1371/journal.pone.0085989> PMID: 24465829
13. Korns MF. Highly Accurate Symbolic Regression with Noisy Training Data. In Genetic Programming Theory and Practice XIII 2016 (pp. 91–115). Springer, Cham.
14. Gandomi AH, Alavi AH, Ryan C, editors. Handbook of genetic programming applications. Switzerland: Springer; 2015 Nov 6.
15. Muttil N, Chau KW. Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*. 2006 Jan 1; 28(3-4):223–38.
16. Bonga JM, von Aderkas P. *In vitro* culture of trees. Springer Science & Business Media; 1992 May 31.
17. Purohit SD, Joshi P, Tak K, Nagori R. Development of high-efficiency micropropagation protocol of an adult tree—*Wrightia tomentosa*. In Plant Biotechnology and Molecular Markers 2004 (pp. 217–227). Springer, Dordrecht.
18. Murashige T, Skoog F. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiologia Plantarum*. 1962 Jul 1; 15(3):473–97.
19. Hastie T, Tibshirani R, Friedman J. Unsupervised learning. In The elements of statistical learning 2009 (pp. 485–585). Springer, New York, NY.
20. Schmidt M, Lipson H. (2013) Eureqa (version 0.98 beta)[software].
21. Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. *Science*. 2009 Apr 3; 324(5923):81–5. <https://doi.org/10.1126/science.1165893> PMID: 19342586
22. Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika*. 1991 Sep 1; 78(3):691–2.
23. Colbourn E, Rowe RC. Neural computing and pharmaceutical formulation. Encyclopedia of Pharmaceutical Technology. New York, USA: Marcel Dekker. 2005:145–57.
24. Faris H, Sheta A. Identification of the Tennessee Eastman chemical process reactor using genetic programming. *International Journal of Advanced Science and Technology*. 2013 Jan; 50:121–40.
25. Klug M, Bagrow JP. Understanding the group dynamics and success of teams. *Royal Society open science*. 2016 Apr 1; 3(4):160007. <https://doi.org/10.1098/rsos.160007> PMID: 27152217
26. Sharma PK, Trivedi R, Purohit SD. Activated Charcoal Improves Rooting in *in vitro*-Derived *Acacia leucophloea* shoots.
27. Mathai MP, Zachariah JC, Samsudeen K, Rema J, Nirmal Babu K, Ravindran PN. Micropropagation of *Cinnamomum verum* (Bercht & Presl.) In Edison S., Ramana KV, Sasikumar B., Babu K, Nirmal and SJ Eapen (eds) Biotechnology of Spices, Medicinal and Aromatic Plants, ISS, IISR, Calicut. 1997:35–8.
28. Chen C. Application of growth models to evaluate the microenvironmental conditions using tissue culture plantlets of *Phalaenopsis* Sogo Yukidian 'V3'. *Scientia Horticulturae*. 2015 Aug 6; 191:25–30.
29. Arab MM, Yadollahi A, Shojaeiyan A, Shokri S, Ghoghah SM. Effects of nutrient media, different cytokinin types and their concentrations on *in vitro* multiplication of G× N15 (hybrid of almond× peach) vegetative rootstock. *Journal of Genetic Engineering and Biotechnology*. 2014 Dec 1; 12(2):81–7.
30. Duffy J, Engle-Warnick J. Using symbolic regression to infer strategies from experimental data. In Evolutionary computation in Economics and Finance 2002 (pp. 61–82). Physica, Heidelberg.
31. Koza JR, Bennett FH, Andre D, Keane MA. Genetic programming III: Darwinian invention and problem solving [Book Review]. *IEEE Transactions on Evolutionary Computation*. 1999 Sep; 3(3):251–3.
32. Purohit SD, Kukda G. *In vitro* propagation of *Wrightia tinctoria*. *Biologia Plantarum*. 1994 Dec 1; 36(4):519–26.
33. Vladislavleva EY. Model-based problem solving through symbolic regression via pareto genetic programming. CentER, Tilburg University; 2008 Aug.