Electronic Journal of Biotechnology 18 (2015) 347–354

HOSTED BY

Contents lists available at ScienceDirect

Electronic Journal of Biotechnology

Soft computing model on genetic diversity and pathotype differentiation of pathogens: A novel approach

Hüseyin Gürüler [a,*], Musa Peker [a], Ömür Baysal [b,*]

[a] Department of Information Systems Engineering, Faculty of Technology, Mugla Sitki Kocman University, 48000 Mugla, Turkey
[b] Department of Molecular Biology and Genetic, Faculty of Life Sciences, Mugla Sitki Kocman University, 48000 Mugla, Turkey

ARTICLE INFO

Article history:
Received 20 March 2015
Accepted 30 June 2015
Available online 22 August 2015

Keywords:
Computational biology
Genetic diversity
Molecular markers
Plant pathogens
Predictive information
Soft computing

ABSTRACT

Background: Identifying and validating biomarkers' scores of polymorphic bands are important for studies related to the molecular diversity of pathogens. Although these validations provide more relevant results, the experiments are very complex and time-consuming. Besides rapid identification of plant pathogens causing disease, assessing genetic diversity and pathotype formation using automated soft computing methods are advantageous in terms of following genetic variation of pathogens on plants. In the present study, artificial neural network (ANN) as a soft computing method was applied to classify plant pathogen types and fungicide susceptibilities using the presence/absence of certain sequence markers as predictive features.
Results: A plant pathogen, causing downy mildew disease on cucurbits was considered as a model microorganism. Significant accuracy was achieved with particle swarm optimization (PSO) trained ANNs.
Conclusions: This pioneer study for estimation of pathogen properties using molecular markers demonstrates that neural networks achieve good performance for the proposed application.

1. Introduction

Biotechnological improvements have provided powerful methods for simultaneously measuring cellular metabolisms under different conditions and periods of expression levels on lots of genes related to the metabolism of the cell [1]. In the era of modern biotechnology, several molecular techniques have been developed for the genetic studies and characterizations of different organisms, among which inter-simple sequence repeat (ISSR), sequence-related amplified polymorphism (SRAP), and simple sequence repeat (SSR) analysis are well established and widely used. However, detection of important and necessary data from these datasets requires long and difficult processes [2]. A key step in the analysis of genetic diversity is to provide detailed information regarding determination groups and their variances in similar expression patterns [3,4]. As an example for microbiological application, automatically operating technology like soft computing has been used for analyzing complex data related to plant pathogens [3]. These technologies are essential for microbiologists in terms of minimizing the workload. In previous studies, reasonably accurate results have been obtained in modeling and estimation in the fields of molecular biology and genetic characterization. Therefore, soft computing methods provide numerous opportunities for bioinformatics by producing especially low-cost and practical solutions [4,5].

In previous works, some soft computing methods have been employed for epidemiologic studies related to plant diseases. For instance Rumpf et al. [6] identified healthy and diseased plants in sugar beet leaves. In the study, a support vector machine (SVM) algorithm was used and a success rate of up to 97% was achieved. Bauer et al. [7] used k-nearest neighbor, Gaussian mixture and conditional random field methods to decompose diseased plants in sugar beet leaves and obtained 86 and 91% success rates, respectively. Li et al. [8] investigated three different leaf diseases by using the methods of principal component analysis and discriminant analysis, where 96.7, 93.3, and 86.7% success rates were obtained respectively. In another study, Luaces et al. [9] favored the SVM to identify the rust disease in coffee plants. As a result of experiments, they obtained 90 and 78% success rates. Romer et al. [10] used the SVM algorithm for identifying the rust disease in wheat leaves. As a result of experiments, they obtained a success rate of 93%. Wang et al. [11] proposed a neural network based model to identify the pathogen named *Phytophthora infestans* that causes destructive disease on tomato. Bravo et al. [12] also investigated the spectral reflectance difference between healthy and rust diseased wheat plants. Obtained results showed accordance with evaluation of data mining and field observations. In addition to early identification of plant diseases, automated methods are also important in terms of assessing the genetic diversity and pathotype formations. To the best our knowledge, studies regarding automated methods on the subject of

* Corresponding authors.
E-mail addresses: hguruler@mu.edu.tr (H. Gürüler), omurbaysal@mu.edu.tr (Ö. Baysal).
Peer review under responsibility of Pontificia Universidad Católica de Valparaíso.

http://dx.doi.org/10.1016/j.ejbt.2015.06.006

pathotype detection and fungicide resistance have not yet been studied. Even though fungicide resistance can be evaluated in agar diffusion leaf disc test using petri plates, estimation of resistance occurrence from genetic differentiation using soft computing techniques is an original and cost effective method.

In the present study, a soft computing model providing predictive information about pathotype diversity of plant pathogens and fungicide resistance was developed and significant accuracy was achieved with particle swarm optimization (PSO) trained artificial neural networks (ANNs).

2. Materials and methods

2.1. Data resources for screening in evaluation and biological validation

Dataset used in this study were constructed by experts from 3 different countries (Israel, Czech and Turkey) in the frame of international collaboration. Data on *Pseudoperonospora cubensis* isolates and their properties related to *mefenoxam* sensitivity testing and molecular diversity studies have been evaluated on 800 isolates of three different countries using SSR, ISSR, and SRAP biomolecular markers, some of which have been published by Polat et al. [13]. Amplified bands from each primer were scored as present (1) or absent (0). With the exception of consistently amplified bands, smeared and weak bands were also scored as (-1) in the analysis. The pairwise genetic distances for phylogenetic relationships among strains were estimated using Nei's coefficient [14]. A dissimilarity matrix was computed and a weighted neighbor-joining tree was generated with Power Marker version 3.25 using the datasets obtained from ISSR and SRAP [15]. A consensus tree was created in NEXUS format for viewing in tree-view [16], the nodes are being supported by bootstrap analysis (1000 replicates) as given in [17].

Additional statistics were computed to estimate the grade of polymorphism among the studied isolates. The percentage of polymorphic loci, Shannon's Information index, and the Nei's gene diversity within the collection analyzed were calculated using POPGENE, version 1.31 [18].

2.2. Artificial neural network

ANNs are mathematical systems that consist of many processing units weighted and connected to each other [19]. This processing unit receives signals from other neurons which it combines and transforms to reveal a numerical result. In general, the processing units roughly correspond to the actual neurons and interconnected in a network; this structure constitutes neural networks. In this study, a multilayer perceptron (MLP) network model was used. Basically, there are three layers in this type of networks; the input layer that holds data entering neural network, hidden layer or layers that educate themselves according to the desired result, and finally an output layer which presents output values.

2.3. Particle swarm optimization

In PSO, each solution is called as particle in the search-space. All particles have relevancy value evaluated by the relevancy function to be optimized and particle velocity information directing their movements. Particles follow the existing optimum particles in the search-space [20].

PSO is initialized with random particle swarm and the optimum value is iteratively searched. In each iteration, each particle is updated according to the best two values. The particle with the optimal relevancy value is assigned the notation *pbest*. This value is noted for later use. The second best value is the best relevancy value found by any particle in swarm called *gbest*. It is the best global value in the swarm [21,22].

Swarm matrix with D swarm dimension and n particle size is described in [Equation 1] as follows.

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nD} \end{bmatrix}_{nxD.} \quad \text{[Equation 1]}$$

According to the swarm matrix ith particle is described in [Equation 2] as:

$$x_i = [x_{i1},\ x_{i2},\ x_{i3},\ \ldots, x_{iD}] \quad \text{[Equation 2]}$$

and the *pbest*, best relevancy value found by the particle so far, is

$$pbest_i = [p_{i1},\ p_{i2},\ p_{i3},\ \ldots,\ p_{iD}] \quad \text{[Equation 3]}$$

gbest within the population

$$gbest = [p_1,\ p_2,\ p_3,\ \ldots,\ p_D]. \quad \text{[Equation 4]}$$

Fig. 1 shows the velocity and position updating of a particle.

ith is described as a velocity vector indicating the amount of change in each position of the particle.

$$v_i = [v_{i1},\ v_{i2},\ v_{i3},\ \ldots,\ v_{iD}]. \quad \text{[Equation 5]}$$

Particle's velocity and position are updated according to the following equations, respectively.

$$v_i^{k+1} = v_i^k + c_1.rand_1^k.\left(pbest_i^k - x_i^k\right) + c_2.rand_2^k.\left(gbest_i^k - x_i^k\right) x_i^{k+1} = x_i^k + v_i^{k+1} \quad \text{[Equation 6]}$$

where k denotes the number of iterations and i the number of particles. If the particle swarm matrix consists of n rows, it means that ith line is being mentioned. c_1 and c_2 values which are the learning factors, pull the particle to *pbest* and *gbest* values. c_1 and c_2 are usually selected as equal and in {0, 4} range. c_1 allows particle to move according to the particle's own experience, and furthermore c_1 allows particle to move according to the experience of other particles in the swarm.

2.4. The soft computing model: PSO-based MLP network

In this study, unlike classical training algorithms, the PSO, a powerful optimization algorithm, was preferred for weight adjustments of

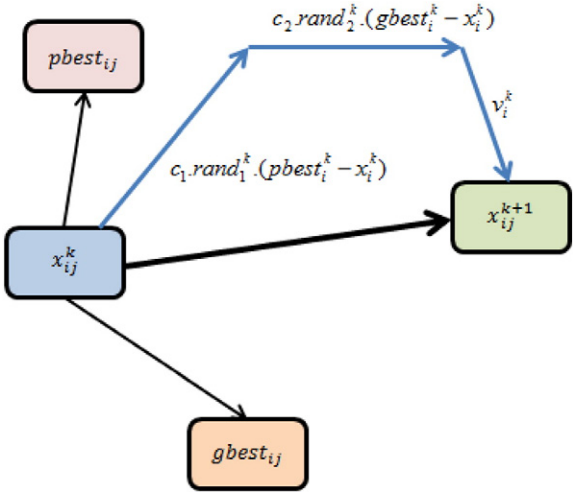

Fig. 1. The velocity and position updating of a particle at kth generation [21].

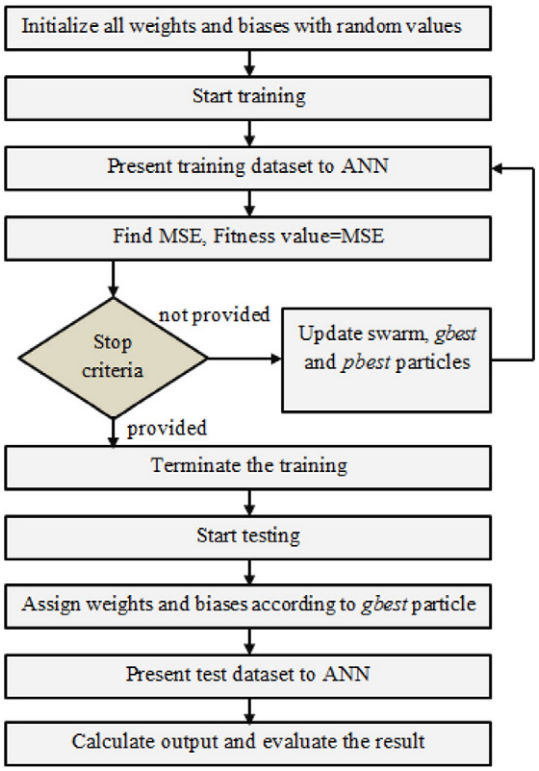

Fig. 2. Flow chart for training and testing of PSO-based network.

network. Hence, the realization of learning in ANNs weight values between the layers should be appropriately updated.

In Fig. 2, a flow chart in which testing and training of network with PSO is presented. In learning phase, primarily, weights holding the numerical value of connections between layers take random values. These weight values represent particle values for PSO. The number of connections between the layers denotes the size of particles [21].

The network is established according to each particle and training examples are sent to the network respectively [23]. After all the samples are submitted to the network, mean squared error (MSE) is calculated and the obtained value is regarded as the particle's relevancy value. Fundamentally, an error is the difference between an output vector and its target vector. This relevancy value is assigned as *pbest* value of the particle; the best relevancy value among the particles is assigned as the *gbest* value.

If relevancy value (error) is not at an acceptable level, particles are updated with *pbest* and *gbest* values. The network is re-established according to the new particle values, examples are given to the network again and the relevancy value calculation is performed. These processes continue until the best relevancy value is obtained (*gbest*) and reaches to the desired value or the maximum iteration [21].

When the error is reduced to acceptable level, the testing process begins. This time, network is established according to the *gbest* particle values. Test samples are sent respectively to the input layer of the network and the resulting values are given as output of the example. If any threshold is not applied to the output of the network, the last obtained *gbest* value gives the classification performance of the network.

The preferred neural network structure in this study is shown in Fig. 3. Here, 68 attributes obtained from SSR, ISSR, and SRAP sequences in feature extraction stage were presented as an input to the neural network structure. Output consists of 5 values for type of pathogens (O1) and output consists of 4 values for fungicide resistance (O2).

2.5. The other soft computing algorithms used in the study

In this study some soft computing algorithms have been used to realize classification. ANNs are weighted mathematical system consisting of many neurons and layers, which are connected to each other. In this study, the MLP neural network model was used as mentioned above. The research was performed on five different algorithms as an alternative to the ANN. SVM is a method of classification and regression classes that can be easily used on normally difficult to be classified in basic (linear or nonlinear) datasets with the help of its core functions [24]. Logistic regression measures the relationship between categorical dependent variable and continuous independent variable(s) in terms of probability [25]. The k-nearest neighbor (kNN) algorithm is an instance-based, a non-parametric and the simplest of all machine learning algorithms that store all available cases and classify new cases based on a similarity measure refer to distance [26,27]. Naïve Bayes (NB) is a well known statistical learning algorithm. NB is a simple probabilistic classifier that is highly scalable, requiring a number of parameters linear in a learning problem [28]. Random Forest uses multiple decision trees during the classification process to obtain more accurate results. Therefore, Breiman [29], proposed the unification of the

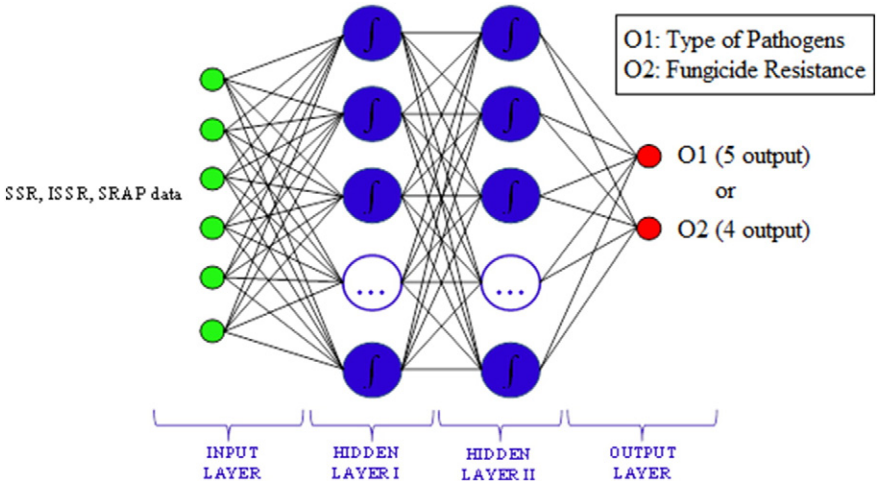

Fig. 3. Neural network architecture.

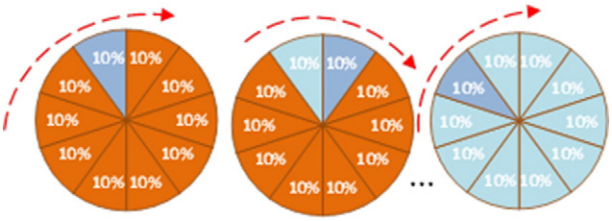

Fig. 4. Ten-fold cross-validation.

multivariate decision tree each trained with a large number of different education clusters instead of producing a single decision tree.

2.6. Evaluation methods

Five different evaluation criteria were used: accuracy, sensitivity, specificity of the classification, MSE, and k-fold cross-validation.

Classification accuracy (CA): Classification accuracy is widely used as a metric for evaluation of machine learning systems. The classification accuracy is defined as the percentage of test data that can be correctly classified [Equation 7]:

$$CA = \frac{Correct\ Classified\ Patterns}{Total\ Patterns} \times (100\%). \quad \text{[Equation 7]}$$

Sensitivity and specificity: sensitivity measures the percentage of actual positives which are correctly identified. Specificity measures the percentage of negatives which are correctly identified. The following expressions for the sensitivity and specificity analyses were used:

$$sensitivity = \frac{TP}{TP + FN}(\%) \quad \text{[Equation 8]}$$

$$specificity = \frac{TN}{FP + TN}(\%). \quad \text{[Equation 9]}$$

Here, TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative, respectively.

MSE: to evaluate the accuracy of the model with a different way, the MSE criterion was also computed, see [Equation 10]. Basically, the ANN model achieves a better performance when MSE is small.

$$MSE = \frac{\sum_{j=0}^{P}\sum_{i=0}^{N}\left(t_{ij}\text{-}y_{ij}\right)^2}{NP} \quad \text{[Equation 10]}$$

where P is the number of output possessing elements and N is the number of exemplars in the dataset. t_{ij} and y_{ij} represent target output and obtained network outputs, respectively.

k-fold cross-validation: The dataset is divided into k groups randomly. The first group is reserved for the test. The model is established with the remaining groups. The established model is estimated on the data which reserved for the test and the accuracy rate is calculated. The process is repeated k times and the model's accuracy rate is the average of k accuracy rates. In the present study, ten-fold cross-validation approach [30,31] has been used to estimate the performance of classifiers as suggested optimal number of folds (Fig. 4).

Regression coefficient: regression analysis is used in order to determine the relationship between two or more variables that have cause–effect relationship between them and to make forecasts or predictions regarding that subject using these relations. Where the regression value is close to 1, the linear dependence between X and Y variables is strengthened.

3. Results and discussion

Appropriate architecture was assessed after several attempts to classify pathotypes and fungicides resistivity. Neural network architectures were identified as *I–H1–H2–O*. Where *I* represents the number of neurons in the input layer. Input values were obtained as a result of ISSR, SSR, and SRAP sequences. These input values are shown in Table 1. *H1* and *H2* represent the number of hidden neurons for layer 1 and layer 2, respectively. *O* refers to the number of neurons in the output layer.

In this study, gel scores of molecular genetic markers SSR, ISSR, and SRAP sequences were used as input values. Output values were determined as pathogen type and fungicide susceptibility. Output values were intended to identify the type of pathotype and resistivity value. In order to use these values in neural network, an encoding has been developed. The pathotypes are encoded as SW, 3, 5, 6, and 7. The resistance values were encoded as SW, P, R, and S. SW represents smeared and weak input values detected during analysis. R represents positive resistance, S represents negative resistance and P represents (-). SW values were coded as -1. Table 2, shows the details about this encoding. For instance; suppose that the *O1* output is 3. In this case, the encoding of *O1* is {0 1 0 0 0}. This means that only the selected output value is set to 1 and the others take the value 0. Likewise, suppose that the *O2* output is -1. In this case, the encoding of *O2* is {1 0 0 0} as emphasized in Table 2.

Experiments to assess both resistance and pathotype were performed at three stages, depending on whether or not using the data contained SW.

3.1. Case study 1

At the stage of experiment 1, all collected data were used. These data include SW values. The number of samples used in the experiments is 800. The architecture used for the resistance detection is 68-10-5-5. At the stage of this experiment, the architecture used for the separation of pathotype is 68-15-10-4. These values were preferred for obtaining good results in terms of similarity between the experiment results and

Table 1
Input values obtained from gel scores of molecular genetic markers ISSR, SSR, and SRAP sequences.

	ISSR marker			SRAP marker			SSR marker		
Line no.	112 1500	112 390	808 1100	me2em14 550	me4em14 500	me8em 14 750	TAA52 1200	TAA52 650	B14B03 350
1	0	0	1	-1	-1	1	0	0	0
2	1	0	1	1	0	1	0	0	0
3	0	0	1	1	0	1	0	1	1
4	0	0	1	1	0	0	0	0	0
5	1	0	1	1	0	1	0	0	1
6	1	0	0	-1	0	0	0	0	0
7	1	0	1	0	0	1	1	0	0
8	0	0	0	-1	-1	0	-1	-1	-1
800	0	0	1	0	1	0	0	0	0

Table 2
Encoded output values.

O1	*O2*	O1					O2			
		-1	3	5	6	7	-1	0	1	2
3	1	0	1	0	0	0	0	0	1	0
3	1	0	1	0	0	0	0	0	1	0
3	1	0	1	0	0	0	0	0	1	0
6	2	0	0	0	1	0	0	0	0	1
6	-1	0	0	0	1	0	1	0	0	0
6	-1	0	0	0	1	0	1	0	0	0
6	1	0	0	0	1	0	0	0	1	0
7	1	0	0	0	0	1	0	0	1	0
7	-1	0	0	0	0	1	1	0	0	0
6	1	0	0	0	1	0	0	0	1	0
3	0	0	1	0	0	0	0	1	0	0
3	2	0	1	0	0	0	0	0	0	1

obtained in biological assays. The Iteration-Error results obtained are presented graphically in Fig. 5a and Fig. 6a.

Regression graphs, which depend on the actual output and the expected output values, are shown in Fig. 7a and Fig. 8a. As shown in these figures, the regression value is around 0.82.

The statistical results obtained at this stage are shown in Table 3. Despite the SW values in the dataset, a success rate of over 85% for two classification problem was obtained. The best result in pathogen detection was obtained with 655th iteration. The best result in detecting resistance was obtained with 635th iteration.

3.2. Case study 2

By this stage of the experiment, the data having input values that contain SW had been eliminated. Therefore, the number of samples used in the experiments is 680. The architecture used for the resistance detection was 68-10-5 and the architecture used for the detection of pathogen was 68-15-4. The experiment results were similar in assays of biological experiments. The Iteration-Error results obtained are presented graphically in Fig. 5b and Fig. 6b.

Regression graphs are shown in Fig. 7b and Fig. 8b. As shown in these figures, the regression value is around 0.9 for both classification problems.

The statistical results obtained at this stage are shown in Table 3. Despite the SW values in the input data, a success rate of over 95% for two classification problem was obtained. The best result for pathogen detection was obtained with 576th iteration. The best result in detecting resistance was obtained with 555th iteration.

3.3. Case study 3

At the stage of this experiment, the data containing SW values in input or output values had been eliminated. The number of samples used in the experiments is 360. The architecture used for the resistance detection was 68-15-5. The architecture used for the detection of pathogen was 68-10-4. The experiment results were similar to the ones obtained in biological assays. The Iteration-Error results obtained are presented graphically in Fig. 5c and Fig. 6c.

Regression graphs are shown in Fig. 7c and Fig. 8c. As shown in these figures, the regression value was around 0.95.

The statistical results obtained at this stage are shown in Table 3. Significant increase in the success rate was observed after eliminating all the data containing SW, where a success rate of over 98% was achieved. Best result in pathogen detection was obtained with 465th iteration. The best result in detecting resistance was obtained with the 427th iteration.

3.4. Comparison analysis on ANN tools

The tests were carried out on the most widely used five different soft computing tools. These were Matlab, WEKA, Orange, Knime, and EasyNN. Obtaining similar results have shown to be platform-independent. Table 4 shows the ANN results obtained from the mentioned soft computing tools and specific ANN software.

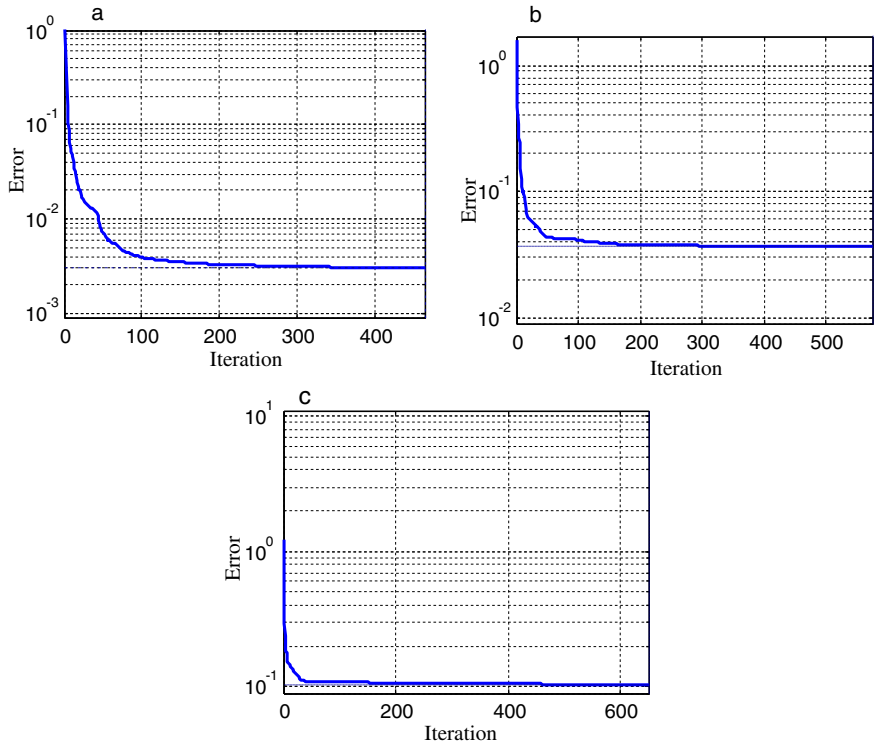

Fig. 5. Iteration-Error graphs for the pathotype differentiation a) for experiment 1, b) for experiment 2, c) for experiment 3.

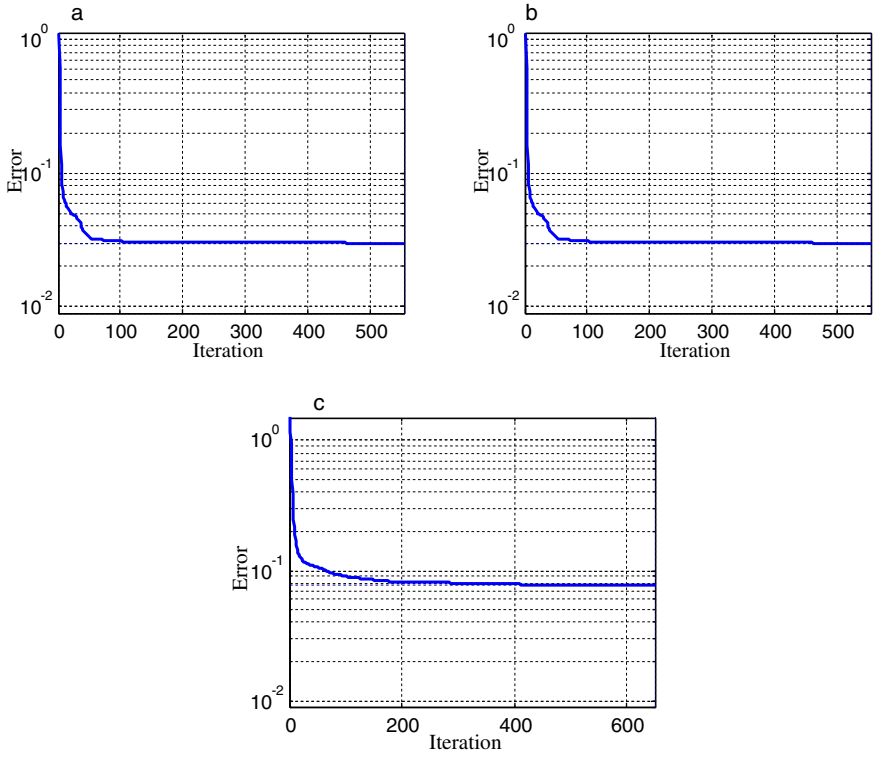

Fig. 6. Iteration-Error graphs for fungicide resistance a) for experiment 1, b) for experiment 2, c) for experiment 3.

3.5. Comparison analysis on classification algorithms

In this study, different classification algorithms were also applied to determine the most effective classification algorithm. These algorithms are ANN-BP, SVM, Logistic regression, kNN, Naive Bayes, and Random Forest algorithms, respectively. Obtained results are presented in Table 5.

When examining Table 4, ANN seems to offer the best solution. High accurate results are obtained with Logistic regression and SVM as well. The lowest accuracy rate was obtained with the KNN and Naive Bayes algorithms. Therefore, in the latter part of the study, ANN optimization has been performed since it provides the best accuracy value. Thus, 98% success was gained by implementation of PSO with improving optimization on ANN training instead of standard training (back

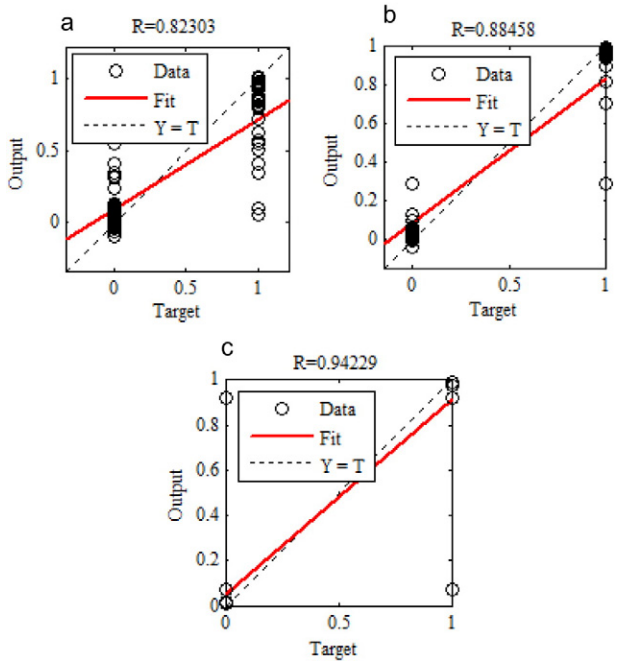

Fig. 7. Regression graphs for pathotype differentiation a) for experiment 1, b) for experiment 2, c) for experiment 3.

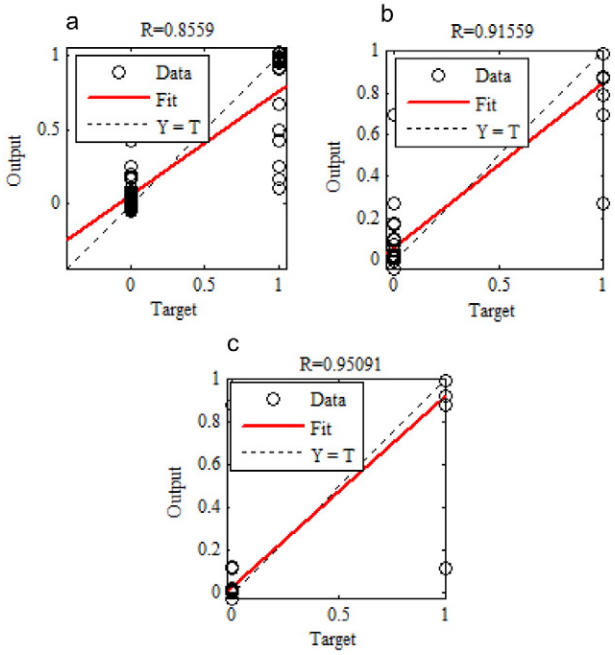

Fig. 8. Regression graphs for fungicide resistance a) for experiment 1, b) for experiment 2, c) for experiment 3.

Table 3
The results obtained from performance evaluation criteria.

Experiments	Output name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Iteration	Regression value	MSE
Exp.-1	Pathotype differentiation	85.4	88.5	91.2	655	0.82	0.1032
	Fungicide resistance	86.7	90.1	92.4	635	0.85	0.0772
Exp.-2	Pathotype differentiation	95	96.5	97.05	576	0.88	0.03665
	Fungicide resistance	95.7	97.8	98.5	555	0.91	0.02995
Exp.-3	Pathotype differentiation	98.2	99.3	99.8	465	0.94	0.00299
	Fungicide resistance	98.5	99.6	100	427	0.95	0.00857

propagation). PSO is a powerful and widely used optimizing algorithm. Due to the constraints of training structures, PSO trained ANN was conducted in Matlab.

The originality of the manuscript is introducing the pioneer research that using soft computing methods with molecular markers used for genetical discrimination of plant pathogens. Adapting ANN on new molecular markers should be considered as a future work. Because molecular markers showing high polymorphism is required a new study on new pathogens after detailed screening according to sequence of markers. To the best of our knowledge there is no open source alternative dataset that is similar to be compared. Therefore there is no possibility for comparing results of different datasets with another published study in which ISSR and SRAP markers have been used in purpose of pathogenic properties and chemical resistance with ANN system. Therefore, further studies, which use sequence data to be provided from specific gene encoding proteins are required to improve the classification properties on pathogens.

We have found a few partially similar studies obtained through intensive scans. Ornella and Tapia [32] proposed supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. Gene expression profiles have been used to predict mandarin clementine varieties (*Citrus clementina* Hort. ex Tan.) by means of two independent supervised learning algorithms: SVMs and prediction analysis of microarrays [33]. This study has also pointed that the small genetic variability existing among these varieties makes molecular markers ineffective in distinguishing genotypes within a particular species. The tool so called ISSR-PCR, which use self-organizing maps as soft computing was developed for discrimination and genetic structure analysis of *Plutella xylostella* populations native to different geographical areas. The classification methods have given results with less than 1.3% of misclassified individuals [34]. In the other study, different bioinformatics algorithms such as SVM and Naive Bayes have been used to identify cultivars of olive trees based on RAPD and ISSR genetic marker datasets generated from PCR reactions. The results showed that data mining techniques can be effectively used to distinguish between plant cultivars [35]. In order to investigate the genetic diversity of *Ligula intestinalis* populations, nine ISSR markers were applied to populations from nine geographical areas around the world and ten host species. Major genetic differentiation was found to be correlated to five broad geographical regions (Europe, China, Canada, Australia, and Algeria). SOMs are considered to provide an efficient alternative tool for mapping the genetic structures of parasite populations [36].

Table 4
The comparison of NN results.

Experiments	Output name	Matlab	Weka	Orange	Knime	EasyNN
Exp.-1	Pathotype differentiation	80.07	80.06	80.05	80.1	80.02
	Fungicide resistance	81.14	81.15	81.12	81.13	81.10
Exp.-2	Pathotype differentiation	90.30	90.22	90.25	90.25	90.24
	Fungicide resistance	91.12	91.03	91.10	91.07	91.03
Exp.-3	Pathotype differentiation	94.48	94.30	94.35	94.50	94.45
	Fungicide resistance	94.33	94.15	94.21	94.13	94.12

With this aspect, the presented methodology in this manuscript can confidently be used in different fields of molecular biology and genetics. It should be used in formation of different database considering different properties according to target that is not only in plant pathology but also human pathogens including bacteria and fungi.

4. Conclusions

This study presents a soft computing model for classifying plant pathogens and estimating pathotype differentiation with identification of fungicide resistance levels. Significant accuracy was achieved with PSO-based trained ANNs. Experiments to assess both resistance and pathotype were performed at three stages. First, all the data containing SW (smeared and weak input or output values were detected on biomarkers during analysis) around 85% success rate was achieved using raw data of 800 samples. Secondly, input data containing SW were eliminated. At this stage, around 90% success rate was achieved. In the final step, both input and output data containing SW were eliminated, and 98% success rate was also obtained. We conclude that the use of soft computing methods with molecular biomarkers is a sufficiently powerful tool to discover reliable classification of pathotype and fungicide resistance, which may facilitate reducing labor cost and saving time.

In this study, a high correlation was observed with the results based on biological assays and the soft computing methods. Therefore the results show that supervised classification methods may correctly assign blind samples to varieties when both training and test samples are under the same experimental conditions.

Within this study, we showed that feature ranking and biomarker diagnostic would benefit from the integration of information at key points, if the exact molecular markers are selected. Using this

Table 5
Suggested methods and comparison analysis on other classification of algorithms.

Experiments	Method	Classification accuracy (%)	
		Pathotype differentiation	Fungicide resistance
Exp.-1	ANN-BP	80.05	81.13
	SVM	78.02	78.85
	Logistic regression	79.16	79.85
	kNN	72.05	73.45
	Naive Bayes	69.55	69.87
	Random Forest	80.04	81.25
	ANN-PSO	85.40	86.70
Exp.-2	ANN-BP	90.25	91.05
	SVM	88.55	89.05
	Logistic regression	89.95	90.02
	kNN	83.55	83.24
	Naive Bayes	80.05	81.12
	Random Forest	91.12	91.78
	ANN-PSO	95.00	95.70
Exp.-3	ANN-BP	94.32	94.12
	SVM	93.06	93.45
	Logistic regression	93.25	93.56
	kNN	87.56	87.45
	Naive Bayes	84.40	85.09
	Random Forest	93.39	94.05
	ANN-PSO	98.20	98.50

knowledge coming from clinical observations, laboratory experiments or existing literature, we can select the optimal sequencing measure for a given set of gene identification. Using the optimal measure for sequencing and identification of new biomarkers reduces the number of false positive and false negative results, increases the number of true results, thus reducing the time required for verification and increases the overall efficiency of the process. We hope that the proposed method would influence the biomarker diagnostic applications and will enhance the effectiveness of resulted clinical practices. In addition, it can possibly be used in the agriculture systems in a cost-effective, labor efficient, and time saving way.

Conflict of interest

The authors declare that there is no conflict of interest.

References

[1] Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol 2004;6:281–97. http://dx.doi.org/10.1089/106652799318274.

[2] Mei Z, Zhang C, Khan A, Zhu Y, Tania M, Luo P, et al. Efficiency of improved RAPD and ISSR markers in assessing genetic diversity and relationships in *Angelica sinensis* (Oliv.) Diels varieties of China. Electron J Biotechnol 2015;18:96–102. http://dx.doi.org/10.1016/j.ejbt.2014.12.006.

[3] Mucherino A, Papajorgji P, Pardalos PM. A survey of data mining techniques applied to agriculture. Oper Res 2009;9:121–40. http://dx.doi.org/10.1007/s12351-009-0054-6.

[4] Torres-Avilés F, Romeo JS, López-Kleine L. Data mining and influential analysis of gene expression data for plant resistance gene identification in tomato (*Solanum lycopersicum*). Electron J Biotechnol 2014;17:79–82. http://dx.doi.org/10.1016/j.ejbt.2014.01.003.

[5] Papadimitriou S, Likothanassis SD. Kernel-based self-organized maps trained with supervised bias for gene expression data analysis. J Bioinform Comput Biol 2004;1: 647–80. http://dx.doi.org/10.1142/S021972000400034X.

[6] Rumpf T, Mahlein AK, Steiner U, Oerke EC, Dehne HW, Plümer L. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. Comput Electron Agric 2010;74:91–9. http://dx.doi.org/10.1016/j.compag.2010.06.009.

[7] Bauer SD, Korc F, Forstner W. The potential of automatic methods of classification to identify leaf diseases from multispectral images. Precis Agric 2011;12:361–77. http://dx.doi.org/10.1007/s11119-011-9217-6.

[8] Li J, Gao L, Shen Z. Extraction and analysis of digital images feature of three kinds of wheat diseases. 3rd International Congress on Image and Signal Processing (CISP). Yantai (China), 6. ; 2010. p. 2543–8. http://dx.doi.org/10.1109/CISP.2010.5646912.

[9] Luaces O, Rodrigues LHA, Meira CAA, Bahamonde A. Using nondeterministic learners to alert on coffee rust disease. Expert Syst Appl 2011;38:14276–83. http://dx.doi.org/10.1016/j.eswa.2011.05.003.

[10] Romer C, Burling K, Hunsche M, Rumpf T, Noga G, Plumer L. Robust fitting of fluorescence spectra for pre-symptomatic wheat leaf rust detection with support vector machines. Comput Electron Agric 2011;79:180–8. http://dx.doi.org/10.1016/j.compag.2011.09.011.

[11] Wang X, Zhang M, Zhu J, Geng S. Spectral prediction of *Phytophthora infestans* infection on tomatoes using artificial neural network (ANN). Int J Remote Sens 2008;29: 1693–706. http://dx.doi.org/10.1080/01431160701281007.

[12] Bravo C, Moshou D, West J, McCartney A, Ramon H. Early disease detection in wheat fields using spectral reflectance. Biosyst Eng 2003;84:137–45. http://dx.doi.org/10.1016/S1537-5110(02)00269-6.

[13] Polat I, Baysal O, Mercati F, Kitner M, Cohen Y, Lebeda A, et al. Characterization of *Pseudoperonospora cubensis* isolates from Europe and Asia using ISSR and SRAP molecular markers. Eur J Plant Pathol 2014;139:641–53. http://dx.doi.org/10.1007/s10658-014-0420-y.

[14] Nei M, Chesser RK. Estimation of fixation indices and gene diversities. Ann Hum Genet 1983;47:253–9. http://dx.doi.org/10.1111/j.1469-1809.1983.tb00993.x.

[15] Liu K, Muse SV. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 2005;21:2128–9. http://dx.doi.org/10.1093/bioinformatics/bti282.

[16] Page RDM. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 1996;12:357–8. http://dx.doi.org/10.1093/bioinformatics/12.4.357.

[17] Baysal O, Siragusa M, Ikten H, Polat I, Gumrukcu E, Yigit F, et al. *Fusarium oxysporum* f. sp. *lycopersici* races and their genetic discrimination by molecular markers in West Mediterranean region of Turkey. Physiol Mol Plant Pathol 2009;74:68–75. http://dx.doi.org/10.1016/j.pmpp.2009.09.008.

[18] Yeh F, Yang R, Boyle T. Popgene: Microsoft windows-based freeware for population genetic analysis version 1.32. Alberta, Canada: University of Alberta. Center for International Forestry Research (CIFOR); 1999.

[19] Xingui HE, Shaohua XU. Process neural networks: Theory and applications. Berlin Heidelberg: Springer-Verlag; 2010.

[20] Unler A, Murat A. A discrete particle swarm optimization method for feature selection in binary classification problems. Eur J Oper Res 2010;206:528–39. http://dx.doi.org/10.1016/j.ejor.2010.02.032.

[21] Yalcin N, Tezel G, Karakuzu C. Epilepsy diagnosis using artificial neural network learned by PSO. Turk J Electr Eng Comput Sci 2013;23:421–32. http://dx.doi.org/10.3906/elk-1212-151.

[22] Cao J, Lu H, Wang W, Wang J. A novel five-category loan-risk evaluation model using multiclass LS-SVM by PSO. Int J Inf Technol Decis Mak 2012;11:857–74. http://dx.doi.org/10.1142/S021962201250023X.

[23] Yalcin N. Heuristic algorithm basis artificial neural networks for epilepsy detection. M.Sc. Thesis University of Selçuk; 2012.

[24] Vapnik VN. Statistical learning theory. New York, USA: John Wiley & Sons, Inc.; 1998.

[25] Freedman DA. Statistical models: Theory and practice. Cambridge University Press; 2009.

[26] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967; 13:21–7. http://dx.doi.org/10.1109/TIT.1967.1053964.

[27] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46:175–85. http://dx.doi.org/10.1080/00031305.1992.10475879.

[28] Mitchell T. Machine learning. New York, USA: McGraw Hill; 1997.

[29] Breiman L. Random forests. Mach Learn 2001;45:5–32. http://dx.doi.org/10.1023/A:1010933404324.

[30] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif Intell Med 2005;34:113–27. http://dx.doi.org/10.1016/j.artmed.2004.07.002.

[31] Adl AA, Qian X, Xu P, Vehik K, Krischer JP. Feature ranking based on synergy networks to identify prognostic markers in DPT-1. EURASIP J Bioinf Syst Biol 2013;12. http://dx.doi.org/10.1186/1687-4153-2013-12.

[32] Ornella L, Tapia E. Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. Comput Electron Agric 2010;74:250–7. http://dx.doi.org/10.1016/j.compag.2010.08.013.

[33] Ancillo G, Gadea J, Forment J, Guerri J, Navarro L. Class prediction of closely related plant varieties using gene expression profiling. J Exp Bot 2007;58:1927–33. http://dx.doi.org/10.1093/jxb/erm054.

[34] Roux O, Gevrey M, Arvanitakis L, Gers C, Bordat D, Legal L. ISSR-PCR: Tool for discrimination and genetic structure analysis of *Plutella xylostella* populations native to different geographical areas. Mol Phylogenet Evol 2007;43:240–50. http://dx.doi.org/10.1016/j.ympev.2006.09.017.

[35] Beiki AH, Saboor S, Ebrahimi M. A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. PLoS One 2012;7: e44164. http://dx.doi.org/10.1371/journal.pone.0044164.

[36] Bouzid W, Lek S, Mace M, Ben Hassine O, Etienne R, Legal L, et al. Genetic diversity of *Ligula intestinalis* (Cestoda:Diphyllobothriidea) based on analysis of inter-simple sequence repeat markers. J Zool Syst Evol Res 2008;46:289–96. http://dx.doi.org/10.1111/j.1439-0469.2008.00471.x.