Send Orders for Reprints to reprints@benthamscience.ae

BENTHAM OPEN

CrossMark

The Open Bioinformatics Journal

Content list available at: www.benthamopen.com/TOBIOIJ/

DOI: 10.2174/1875036201811010038

RESEARCH ARTICLE

Identification of Better Gene Expression Data for Mosquito Species Classification Using Radial Basis Function Network Methodology

J. Satya Eswari[1,*] and Ch. Venkateswarlu[2]

[1]*Deparment of Biotechnology, National Institute of Technology Raipur, Raipur, India*
[2]*Department of Chemical Engineering , BV Raju Institute of Technology, Narsapur 502313, India*

Received: January 19, 2018 Revised: March 06, 2018 Accepted: March 14, 2018

Abstract:

Background:

Investigation in bioinformatics has developed promptly in latest years owing to improvements in sequence excavating techniques. Gene sequences in DNA are supplemented with great extent of information, but the intricacy and complexity of this information causes difficulty in analyzing it by using standard classical methods of classification. In this work, a Radial Basis Function Network (RBFN) methodology with self-network arrangement is presented for identification of mosquito species based on the genetic design content of ITS2 ribosomal DNA sequences.

Methods:

A number of data sequences in varying sizes of different vectors corresponding to *Anopheline, Aedes* and *Culex* genera are used to develop genera specific as well as comprehensive RBFN species identifiers. The recall and generalization ability of the proposed species identifiers are analyzed and further validated through bootstrap validation method. The genera specific RBFN identifiers are found to provide accurate identification of mosquito species of individual genera. However, the comprehensive RBFN model is found to exhibit better species identification ability and can be advantageously used for species identification of more mosquito genera.

Results & Conclusion:

The results demonstrate the usefulness of the RBFN methodology for accurate identification of mosquito species depending on the nucleotide data of ITS2 ribosomal DNA sequences.

Keywords: Mosquitoes, ITS2 Sequences, Modeling, Radial Basis Function Networks, Species Classification, Gene sequence.

1. INTRODUCTION

Exploration of bioinformatics has developed hastily in recent years due to the advances in molecular biology techniques. One of the most vital research interests in computational biology is sequence mining. It is more difficult to extract knowledge hidden in the sequences than to generate biological sequences. Models and data analysis has become a crucial aspect of biological sequence mining. Several tasks related to sequence mining such as pattern discovery, classification, prediction and clustering can be carried out by statistical, neural network or data mining models [1 - 5]. Such techniques can be used to capture the knowledge or patterns in order to predict, classify or analyze the sequence data. However, the phenomenal growth of sequence data has made the database search computationally intensive and even more forbidding. Gene sequences in DNA are augmented with huge amount of information, but the intricacy and complexity of this information makes difficult in analyzing it by using standard classical methods a challenging task.

* Address correspondence to this author at the Deparment of Biotechnology, National Institute of Technology Raipur, Raipur, India; Tel: +918919640943; E-mail: satyaeswarij.bt@nitrr.ac.in

Genetic level analysis for identification of the species in mosquito is very crucial as the species act as a vector for several diseases such as malaria, filarial, Japanese encephalitis and dengue. Among these diseases, burden of malaria is fairly high and results in heavy toll of life. *Anopheles* mosquito species act as a vector for malaria which is widely spread throughout the globe [6 - 8]. Among 3,500 species of mosquitoes, nearly 430 are *Anopheles* in which around 30-40 species act as a vector for malaria. *Aedes* is a genus of mosquito originated in tropical and subtropical zones, but it has spread through human activity to all continents with the exception of Antarctica [9]. The genus of *Aedes* contains over 700 species. Many of the *Aedes* species transmit potentially serious human diseases. *Culex* is a genus of mosquito in which several species serve as vectors of diseases. The genus of *Culex* has widespread geographical distribution [10] and consists of about 767 species. Classical methods of distinguishing species through genetic analyses have reduced dependence on error-prone morphological and anatomical bases of classification [11]. In the taxonomic classification it is quite difficult to distinguish the sibling species that bear similar morphological and anatomical features but differ at the genetic level. Therefore, for quick identification of a mosquito vector, it is always possible to target a conserved sequence at the genetic level, which is pretty species specific and shows substantial difference even among the sibling species. Internal Transcribed Spacer (ITS) region is widely used in taxonomy and molecular phylogenetics [12, 13]. The ITS2 region located between the 5.8S and 28S gene is highly conserved and species specific. This region is commonly used for DNA sequencing in mosquito genera of *Anopheles*, *Culex*, and *Aedes* [14]. The ITS2 region has been proven useful for differentiating between closely related species of mosquitoes [15 - 17]. This region has also been extensively targeted for species classification, phylogenetic and RNA structure-related analysis [18, 19].

More precise and efficient computational tools are needed to extract genetic level information in the gene sequences and to distinguish mosquito species with minimum or no misclassification. Artificial Neural Networks (ANNs), as one of the Artificial Intelligence (AI) methods, play important role in scientific investigations. ANNs have been used as prediction and classification tools for several applications in the field of bioinformatics including protein structure prediction [20, 21], DNA sequence analysis and biological pattern recognition [22 - 24], RNA structure related analysis [13, 16, 19] and prediction of bacterial promoter sequences [25 - 27]. Recently, ANNs are also used for species identification in mosquitoes [28 - 30]. The wide use of ANNs for prediction and identification is due to their ease of training and flexibility to process high amounts of information with good generalization ability. Although, ANN are widely used for classification and identification problems, most of the ANN models aim at finding the network parameters using a fixed network configuration. It may not be possible to train the ANN to reach a desired level of performance if the network does not have enough computational units, or if the learning algorithm fails to find the optimal network parameters. Therefore, automatic configuration of the network while establishing optimal network parameters is extremely useful in classification and identification of problems. In this work, a Radial Basis Function Network (RBFN) methodology with automatic network configuration is presented for the identification of mosquito species based on the genetic pattern information content of ITS2 ribosomal DNA sequences. A number of data sequences in varying sizes of different vectors corresponding to *Anopheline, Aedes* and *Culex* are used to develop genera specific RBFN species identifiers. Further, a comprehensive RBFN species identifier is also presented for mosquito species identification of more genera. The recall and generalization ability of the RBFN species identifiers are analyzed and further validated through bootstrap validation method.

2. DATA SELECTION AND ITS SIGNIFICANCE

Precise identification of the target species of mosquito has direct medical and practical implications, particularly in developing vector control strategies. Among the vector species of mosquito born diseases, the species belonging to the genera of *Anopheles*, *Aedes* and *Culex* are reported to be involved in the transmission of a variety of vector-borne diseases. Prioritizing the need for detecting the species that act as vectors for several diseases, the present work is targeted on the identification of species of mosquitoes of these three genera.

2.1. Data Collection

The ribosomal DNA sequence (ITS2) data of 15 species of *Anopheles* genera, 10 species each of *Culex* and *Aedes* genera collected in fasta format from National Center for Biotechnological Information (NCBI) nucleotide data base (www.ncbi.nlm.nih.gov/), is used for the computational experiment. Each sequence collected from NCBI nucleotide data base has a specific accession number. For example, the ITS2 part of representative sample sequence of *Anopheles saporoi* belonging to the *Anopheles* genera has the NCBI accession number AY425338.1. Similarly, the sample sequence of *Aedes albiradius* belonging to the *Aedes* genera has the NCBI accession number FM211137.1 and the *Culex pipens* belonging to the *Culex* genera has the NCBI accession number AM084683.1. The length of the sequences

varies in the range of 200 to 500.

2.2. Data Significance

Correct vector identification is imperative to design strategies for managing vector-borne diseases. Since many closely related species of mosquitoes are nearly indistinguishable morphologically, it is difficult to identify mosquitoes especially sibling species correctly. As a consequence, DNA-based approaches have gained increasing importance in mosquitoes identification. The ITS2 ribosomal DNA sequences have proven to be useful for differentiating the closely related species of mosquitoes. The ITS2 region, located between the 5.8S and 28S gene is found to be highly conserved and species specific, and the gene sequence data of this region deposited in NCBI data bank is considered for classification and analysis of the mosquito species. Though the identification of the species listed in NCBI data is based on different criteria, the NCBI assigns a unique identifier to each species and the ITS2 region of the rDNA sequences corresponding to the unique identifier is used to define the species.

3. RBFN SPECIES IDENTIFICATION METHODOLOGY

In mosquitoes, because of the interspecific variability and intraspecific homogeneity in spacer sequences, the species cannot be easily distinguished by just looking at the sequence information. The development of an efficient species classifier is required to identify the closely related species of mosquitoes. Thus, an RBFN classifier with its automatic configuration is developed for the identification of mosquito species based on the gene sequence information contained in the conserved structure of the transcribed spacer, ITS2 region. The RBFN classifier presented in this work is used to distinguish and identify the species of *Anopheles*, *Aedes* and *Culex* genera. Radial Basis Function Network (RBFN) has been considered as viable tool for applications in systems modeling and state estimation [31 - 35]. The RBFN architecture is simple and consists of one input layer, one hidden layer and one output layer, the structure of which is shown in Fig. (**1**). The RBFN algorithm along with its automatic configuration is given in Appendix A.

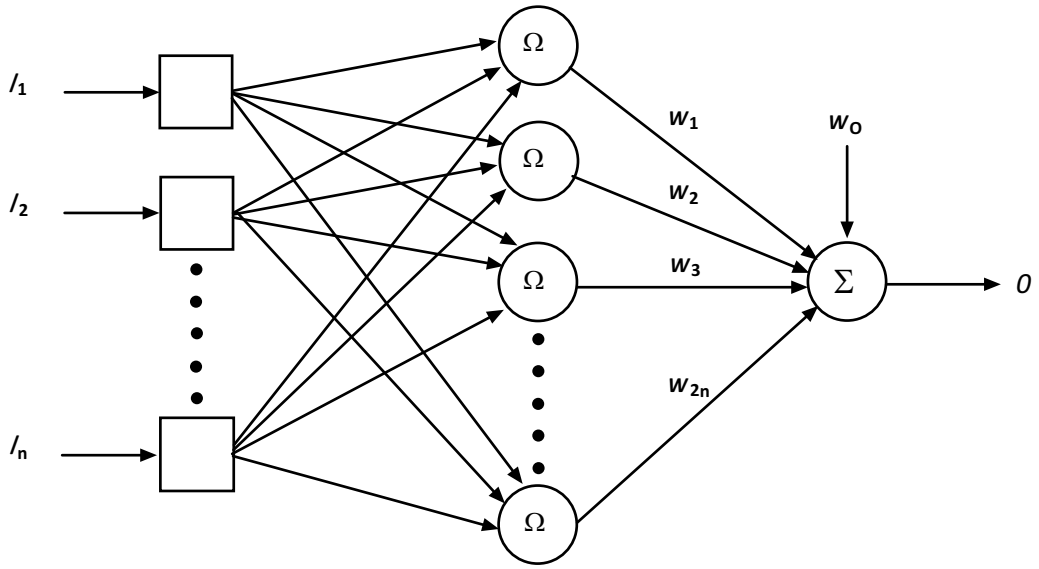

Fig. (1). Radial basis function network.

3.1. Configuring RBFN for Species Identification

In this work, RBFN is configured for mosquito species identification using the data of ITS2 ribosomal DNA sequences corresponding to different mosquito vectors. The automatic configuration of the network is carried out by using a Hierarchically Self-Organizing Learning (HSOL) algorithm. The RBFN used for mosquito species identification has a three layered architecture with input, hidden and output layers, each layer comprising its own nodes, as shown in Fig. (**2**). All nodes in the input layer are connected to the hidden layer nodes, and connections exist between the hidden and output layer nodes. The input layer represents the sequence data, the hidden layer processes the nonlinear information, and the output layer depicts sequence classes. The nodes in the input layer do not perform any numerical processing; all numerical processing is done by the hidden and output layer nodes. The number of nodes in the input

layer is dictated by the chosen sequence encoding schema and the number of the output layer nodes is determined by the species classes representing the network. The nodes in the hidden layer are automatically configured by the HSOL algorithm. The other network parameters that are involved in configuring the RBFN are: the learning rate η, error margin ε_m, initial nominal variance σ_o, initial effective radius r_o, lower bound on radius r_l, radius decrement rate r_d, decay factor for error gradient α, and increment rate for saturation criterion β.

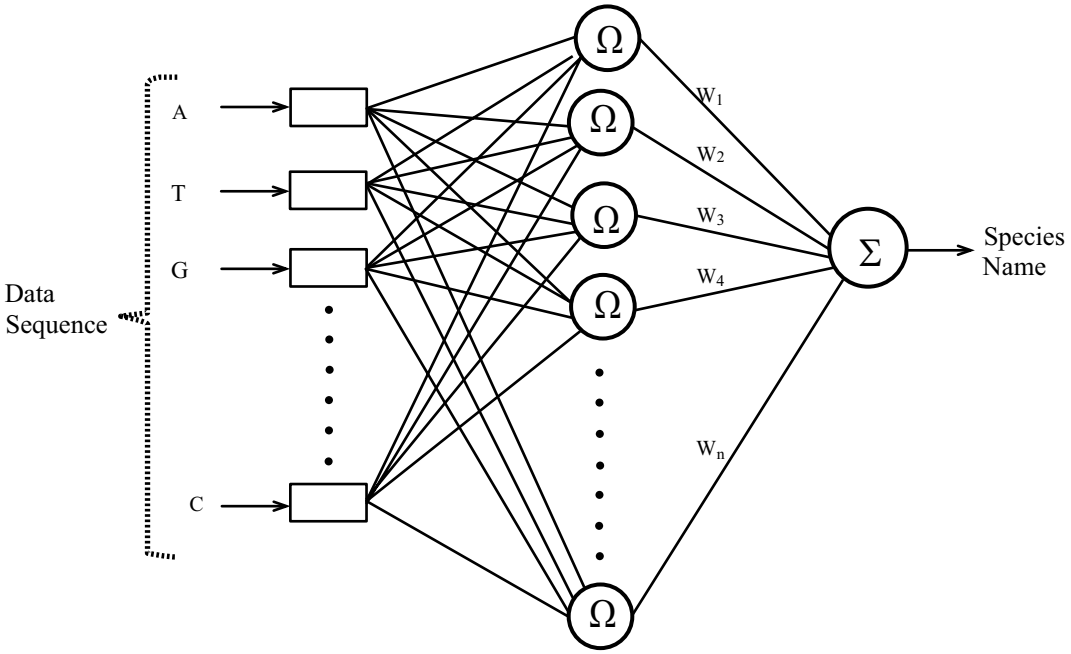

Fig. (2). RBFN structure for genera specific species identification.

3.2. Encoding of Species Data to RBFN

Data encoding plays a crucial role in configuring the RBFN model for species identification. The encoding method can have a significant effect on the learning ability and accuracy of the gene classification system. The data encoding should be in an acceptable format such that the trained network can make clear distinction between the different classes of species data. ITS2 ribosomal DNA sequences of mosquito species are made up of four bases, *Adenine*-A, *Thyamine*-T, *Guanine*-G and *Cytosine*-C. These bases are denoted by a numerical vector which forms the input to the network. The numerical values to encode the input data sequences and to denote the output species are selected so as to yield effective training and classification performance of the network. The bases of ribosomal DNA sequences A, T, G and C representing the inputs to the network are coded with binary values by assigning with A = {0 0}, T = {0 1}, G = {1 0}, C = {1 1}. This input assignment requires the number of nodes in the input layer of the network to be twice the length of the sequence. The output coding that represents the species names is expressed in real coding. The input and output coding assigned for each of the species net is shown in Table **1**. This coding is chosen after testing the network using various input and output coding combinations.

3.3. Genera Specific RBFN Species Identifiers

The RBFN species identifiers are configured for each of the *Anopheles, Aedes and Culex* genera. The data of total 15 species of *Anopheles* genera, 10 species each *Aedes* and *Culex* genera is used for configuring the network. Each species is considered with 4 sequences in which three sequences are used for training and one sequence is used for testing. Thus, *Anopheles* network involves 45 sequences for training and 15 sequences for testing. Each *Aedes* and *Culex* network involves 30 sequences for training and 10 sequences for testing. The bases A, C, T, and G of ribosomal DNA sequences representing the network inputs are coded with binary values and the species names specifying the network output are coded with real numbers. The network structure shown in Fig. (**2**) is employed to treat the data given in Table **1**. The normalized data of outputs is used to train the networks along with the corresponding input data. The network interconnection weights are initialized by assigning random numbers in the range of - 0.2 to 0.2. Each genera net training involves the sequential treatment of the data sets of the input sequences to map with the output coding

specified for the species. The species corresponding to each of the genera are randomly selected for training, and the species that are not involved in training are randomly employed for testing.

The network should accommodate the rDNA data sequences of varying lengths in its training and prediction. For this, the network input nodes are to be chosen so as to suit the sequence of any species irrespective of its size. After studying different options, a simple format is employed to manage the network with the input data sequences of different sizes. According to this format, the number of input nodes to the network is specified in terms of binary coding for the sequence that is larger in size. This format facilitates the network to accommodate the sequences of varying sizes by appropriately filling the empty nodes with the binary codes of the last four bases of the respective sequence. This approach of training and prediction to treat the data of rDNA sequences of different sizes has been proven useful in earlier studies [29, 30].

Table 1. Input and Output data coding for species specific RBFN models.

S. No	Genera Name	Species Name	Input Coding	Output Coding
1	*Anopheles*	*aconitus*	A = {0 0}, T = {0 1}, G = {1 0}, C = {1 1}.	*10*
2		*annulipes*		*20*
3		*bancroftii*		*30*
5		*farauti*		*40*
6		*fluviatilis*		*50*
7		*lesteri*		*60*
8		*maculipennis*		*70*
9		*messeae*		*80*
10		*nuneztovari*		*90*
11		*oswaldoi*		*100*
12		*pullus*		*110*
13		*saperoi*		*120*
14		*sinensis*		*130*
15		*stephensi*		*140*
		subpictus		*150*
1	*Aedes*	*albiradius*	A = {0 0}, T = {0 1}, G = {1 0}, C = {1, 1}.	*10*
2		*ashworthi*		*20*
3		*australis*		*30*
4		*circumluteolus*		*40*
5		*cretinus*		*50*
6		*fontenillei*		*60*
7		*nigropterum*		*70*
8		*palpalae*		*80*
9		*punctor*		*90*
10		*simpsoni*		*100*
1	*Culex*	*erraticus nigripalpus*	A = {0 0}, T = {0 1}, G = {1 0}, C = {1 1}.	*10*
2		*pilosus*		*20*
3		*pipiens*		*30*
4		*pallens*		*40*
5		*pipiens*		*50*
6		*salinarius*		*60*
7		*territans*		*70*
8		*tigripes*		*80*
9		*torrentium*		*90*
10		*tritaeniorhynchus*		*100*

In RBFN model building, the HSOL algorithm enables to provide comprehensive learning with automatic recruitment of RBFs while optimizing the network parameters. Input and output mapping comparison of target and actual values continue until all the data sequences of the training species are learned within an acceptable over all error. During the classification and generalization phase, the trained RBFN itself operates in a feed forward manner.

3.4. Comprehensive RBFN Species Identifier

The comprehensive species identifier is a multi-input and single-output RBFN, in which all the data sequences corresponding to the species of the three genera are treated as inputs to the network along with their corresponding species names as the network outputs. As in individual RBFN, the bases A, T, G, C of ribosomal DNA sequences in binary coding form the network inputs, and their corresponding species names in real coding denote the network outputs. The structure of combined RBFN is similar to that in Fig. (**2**), where the data sequences corresponding species of all the three genera, along with the corresponding output representation given in Table **2**, are sequentially and

iteratively used to train the network. The data of 105 random sequences from the 35 species of the three genera are used for training and the untrained data sets of 35 sequences corresponding the species of all the three genera, are used to evaluate the network predictive performance. The number of input nodes used to accommodate the sequences of varying lengths in comprehensive RBFN is the same as in individual RBFN. The network training is performed using HSOL algorithm. The initialization of the network parameters in combined model is the same as in individual model. All the data sets corresponding to input sequences with their binary coding, and the output species with their numerical coding are sequentially used to train the network model. The network parameters are optimized to establish the relationships between input and output data. A quadratic error function, based on the actual and predicted outputs, forms the objective function, which is minimized through iterative convergence.

Table 2. Input and output data coding for comprehensive RBFN model.

S.No.	Genera name	Species name	Input coding	Output coding
1		*albiradius*	A = {0 0}	10
2		*albopictus*	T = {0 1}	20
3		*ashworthi*	G = {1 0}	30
4		*australis*	C = {1 1}	40
5	*Aedes*	*belleci*		50
6		*cinereus*		60
7		*circumluteolus*		70
8		*cretinus*		80
9		*fontenillei*		90
10		*geminus*		100
11		*albitarsis*		110
12		*Annuples*	A = {0 0}	120
13		*arabiensis*	T = {0 1}	130
14		*atroparves*	G = {1 0}	140
15		*culifacies*	C = {1 1}	150
16		*maculipennis*		160
17		*melanoon*		170
18		*nuneztvan*		180
19		*puulus*		190
20	*Anopheles*	*sachrov*		200
21		*saporoi*		210
22		*sinensis*		220
23		*sundacius*		230
24		*superpictus*		240
25		*erraticus*		250
26		*erythrothorax*	A = {0 0}	260
27		*nigripalpus*	T = {0 1}	270
28	*Culex*	*pilosus*	G = {1 0}	280
29		*pipiens*	C = {1 1}	290
30		*pipienspallens*		300
31		*pipiensquinquefasciatus*		310
32		*restuans*		320
33		*salinarius*		330
34		*tarsalis*		340

4. RESULTS AND DISCUSSION

Individual RBFN models are developed for *anopheles, aedes and culex* genera using the input and output data corresponding to the species of these genera. The presentation of data to RBFN is briefed in the earlier section. The RBFN parameters are set as: learning rate (η) = 0.00025, error margin (ε_m) = 0.001, initial nominal variance (σ_o) = 0.0021, initial effective radius (r_o) = 5.0, lower bound on radius (r_l)= 0, radius decrement rate (r_d) = 0.995, decay factor for error gradient (α) = 1.0 x 10^{-09}, and increment rate for saturation criterion (β) = 1.0 x 10^{-05}. For network training, all the data sets representing the species inputs and outputs are sequentially employed in each iteration and iterations are performed until the network is optimally configured *via* HSOL algorithm. The number of iterations required for convergence, number of RBFs generated, training error evaluated and training time required for each of the networks are given in Table **3**. The trained and learned RBFNs are then subjected to assess their recall and generalization

abilities. In the recall phase, the prediction ability of the trained networks is evaluated by using the same input sequences as used for training. Each of these RBFNs has shown almost 100% recall ability. The generalization performance of the networks is evaluated by using the species sequences that are not involved in training. This is referred here as conventional cross validation. Thus, 15 data sequences corresponding to the species of *anopheles* genera, 10 data sequences corresponding to the species of each *aedes* and *culex* genera, are used to test the generalization ability of the individual RBFN models. The comparison of the model predictions with the actual output species representations in Figs. (**3**-**5**) shows the effective generalization performance of the individual RBFN species identifiers. The generalization ability of the RBFN models is further assessed in terms of Mean Squared Error (*MSE*) and correlation coefficient (R^2) defined in Appendix B. These results in Table **4** further demonstrate the efficiency of the RBFN methodology for accurate identification of mosquito species based on the nucleotide data of ITS2 ribosomal DNA sequences.

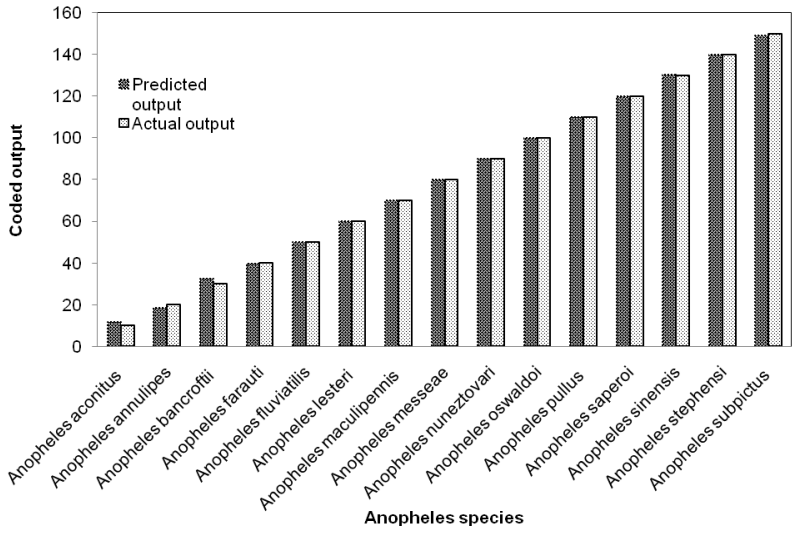

Fig. (3). Comparison of genera specific RBFN model predictions with actual species output codes of *Anopheles* genera.

Table 3. Converged parameters of RBFN species identifiers.

Modeling Configuration	No. of Iterations	No. of RBFNs Generated	Minimum Training Error	Training Time (Seconds)
Anapheles Net	50	28	0.0013	10
Aedes Net	50	26	0.0001	10
Culex Net	50	19	0.0041	10
Combined Net	100	29	0.0015	30

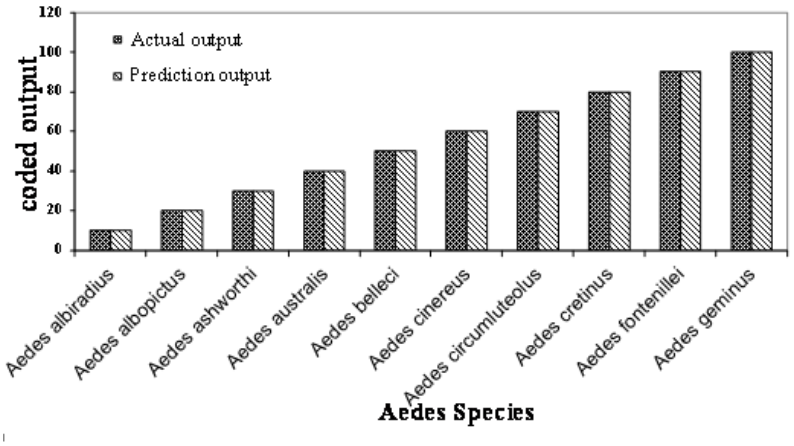

Fig. (4). Comparison of genera specific RBFN model predictions with actual species output codes of *Aedes* genera.

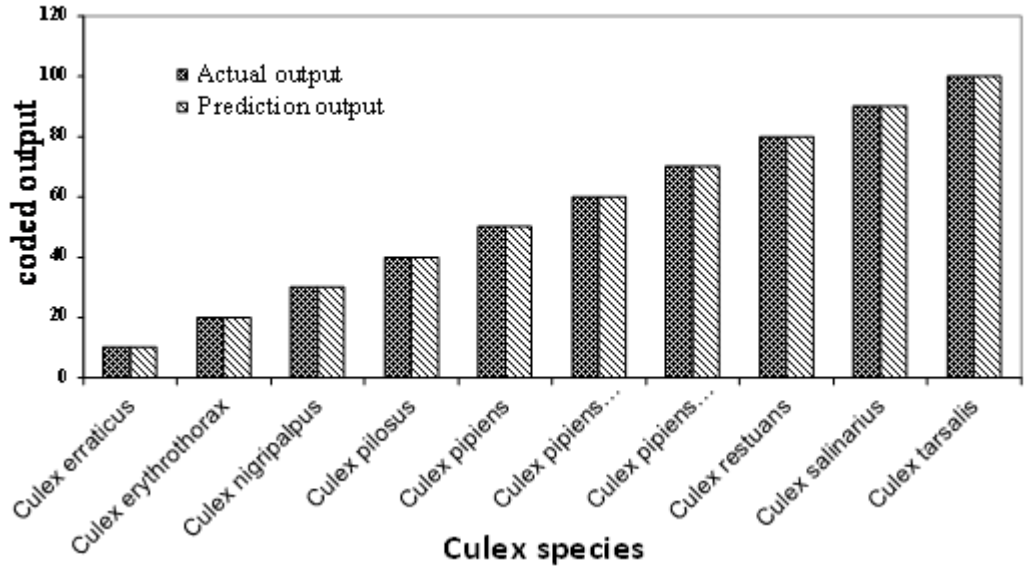

Fig. (5). Comparison of genera specific RBFN model predictions with actual species output codes of *Culex* genera.

Table 4. Predictive performance of RBFN species identifiers.

Species Identifier	Conventional Cross Validation		Bootstrapping Cross Validation	
	MSE	R^2	MSE	R^2
Anapheles net	0.000196	0.98991	0.00323	0.9723
Aedes net	0.000187	0.99437	0.00043	0.9834
Culex net	0.000183	0.99745	0.00421	0.9856
Comprehensive net	0.000283	0.98330	0.00134	0.9753

A comprehensive RBFN model is also developed for species identification of *anopheles, aedes and culex* genera using the input sequence data in binary coding and output species representation in real coding as given in Table **2**. The combined RBFN structure is similar to that of genera specific RBFN structure in Fig. (**2**) with multiple input nodes and single output node. The parameters specified for this RBFN structure are same as in genera specific model structure. The data representation as well as training of this comprehensive network is briefed in the earlier section. In contrast to the genera specific RBFN, this comprehensive RBFN model involves sequential treatment of all the input data sequences representing the species of the three genera along with their respective species output coding. Thus, the data sets in Table **2** are sequentially employed and iterated until the network is optimally configured by HSOL algorithm. The number of iterations required for convergence, number of RBFs generated, training error obtained and training time required for this comprehensive network are given in Table **3**. The entire code was written and executed using C language. The computer configuration used is Intel(R) Core ™ i7-4702MQ CPU @ 2.20GHz 2.20 GHz and the system type is 64-bit Operating System, x64-based processor.

The trained and learned comprehensive RBFN is further studied through conventional cross validation to assess its recall and generalization ability. The comparison of this combined network model predictions with the actual species representations for the recall and generalization phases are shown in Figs. (**6** and **7**). The generalization performance of the combined network is further evaluated in terms of MSE and R^2 as given in last line of Table **4**. These results exhibit better performance of the comprehensive model for species identification of mosquito genera.

Model validation is an important step in ascertaining the reliability of models before they can be used in decision making. The predictive performances of trained and learned RBFNs are further assessed by using bootstrapping procedure. This procedure enables to validate the full *n*-subject model. The RBFN parameters of individual and combined models for this approach are kept same as in earlier model configurations. According to this bootstrap, the *Anopheles* genera net is validated by performing 9 rounds of bootstrapping with the use of all 45 gene sequences. In each round, 40 data sequences corresponding to the species of *Anopheles* genera are randomly assigned for training while the remaining 5 data sequences are assigned for validation. Thus, the model in each round is assessed through

training and validation. The same boot strapping procedure is employed to assess the performance of other genera specific RBFN models and the comprehensive RBFN model. The number of iterations required for convergence, number of RBFs generated, training error obtained and training time required for the genera specific networks and comprehensive network of this bootstrap procedure are found to be nearly same to that of trained and learned RBFN models used in recall and generalization phases. The model validation results of MSE and R^2 evaluated through the bootstrap approach to assess the predictive accuracy of the fitted models are shown in Table **4**. These results confirm the predictive accuracy of the RBFN models developed in this study.

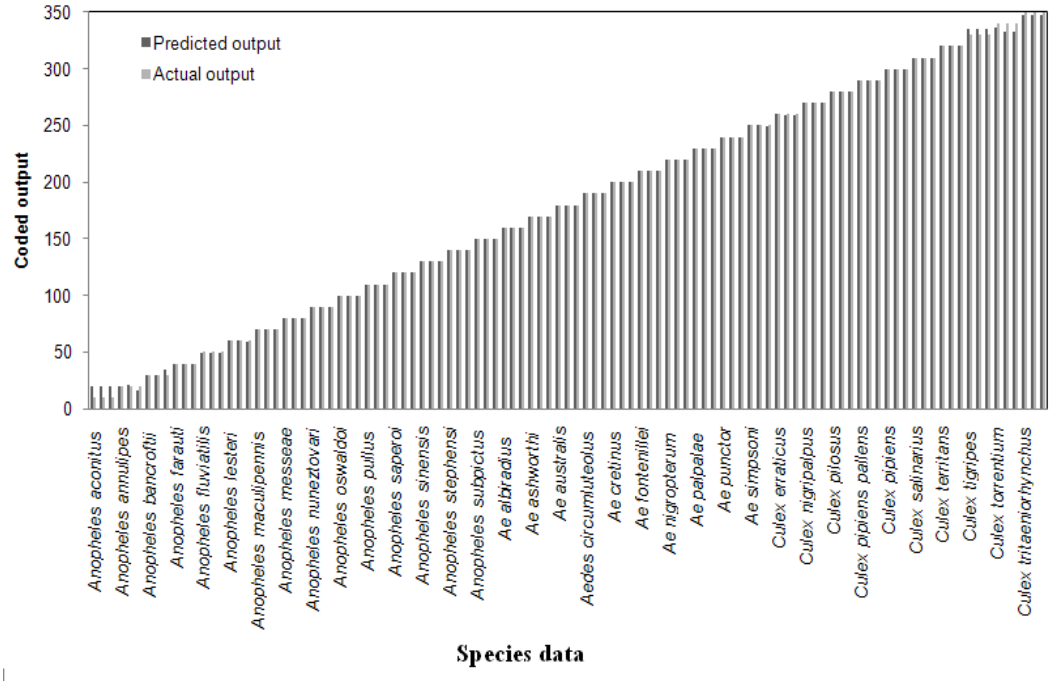

Fig. (6). Comparison of comprehensive RBFN model predictions with actual species output codes in recall phase.

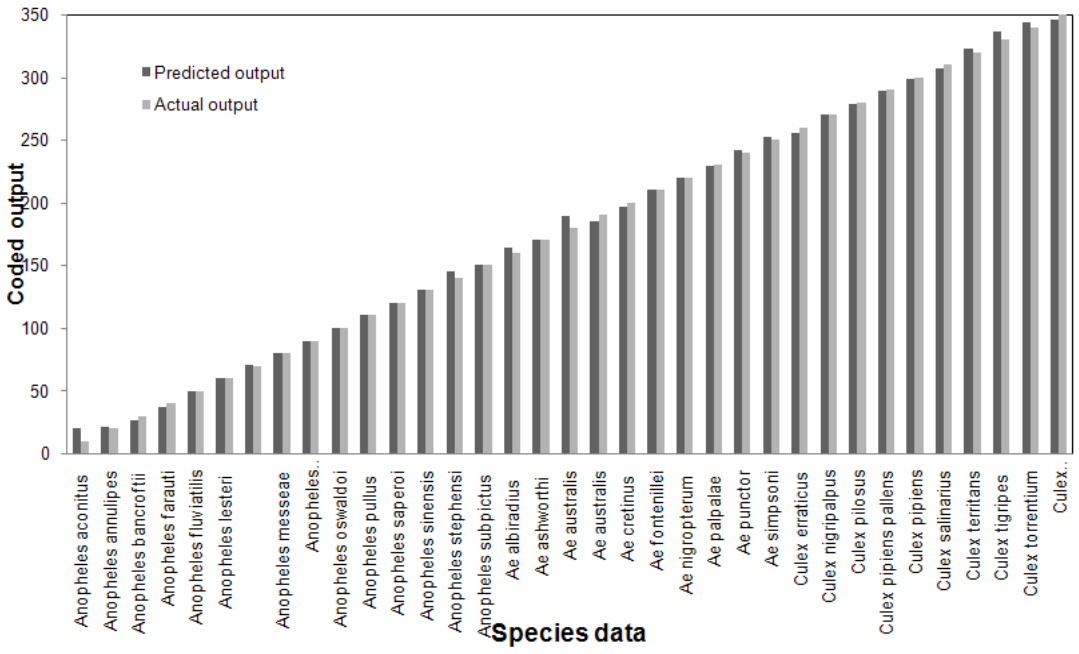

Fig. (7). Comparison of comprehensive RBFN model predictions with actual species output codes in generalization phase.

Recent studies have shown the effectiveness of intelligent computational tools such as Artificial Neural Networks (ANN) for rapid and reliable identification of mosquito species. The genetic data base of mosquitoes has been exploited by such methods to differentiate even closely related species of mosquitoes. Banerjee *et al* (2008) [29] presented a multilayered feed-forward neural network model structure for classification and identification of mosquito species based on the Internal Transcribed Spacer 2 (ITS2) data of ribosomal DNA string. Two network models, namely, a Multi-Input Single-Output Neural Network (MISONN) and Multi-Input Multi-Output Neural Network (MIMONN) have been configured in which the MISONN was found to provide effective performance in distinguishing and identification of mosquito species. However, the above reported work was confined to the analysis of species of single mosquito genera only based on the data of 18 species of *Anopheles* mosquito vectors. The above neural network modeling approach has been further extended by Venkateswarlu *et al*. (2012) [30] for classification and identification of species of more mosquito genera by presenting a Hierarchical Artificial Neural System (HANS) in two levels, in which the first level has a single network that serves as a genera classifier and the second level has multiple networks that perform as species identifiers. Ribosomal DNA sequence data of 34 species of *Anopheles*, *Aedes*, and *Culex* genera, each species with 6 sequences were used for training and validation. The method of HANS was found to provide accurate genera classification and rapid species identification. However, the genera net and species nets involved in HANS were found to require more training efforts. Very recently, Lorenz *et al*. (2015) [36] employed ANN for identification and classification of mosquitoes based on the wing shape characters of the species. A data base of 32 principal components representing the wing shape characteristics of 17 mosquito species of *Anopheles, Aedes* and *Culex* genera were analyzed by using a multi layer perception ANN classifier. Species identification by this method was found to be accurate enough. Although the ANN technique of Lorenzo *et al*. can handle large amount of data, unique variations in wing shape data is required for correct analysis of mosquitoes. Though the above reported ANN configurations are found effective for classification and identification of mosquito species, their major limitation is the heuristic configuration of the network parameters and computational units which may restrict them in achieving the desired performance. The RBFN methodology of this work overcomes such a limitation by automatically configuring the network with optimal selection of computational units.

This work differs from the earlier reported works with respect species involved for analysis as well as the method of species identification. The analysis of the results exhibits accurate identification of mosquito species by genera specific RBFN models, but their application is limited to the identification of species of individual genera. Even though, the genera specific RBFN species identifiers of this work provides marginally improved performance over the Multi-Input Single-Output Neural Networks (MISONN) of earlier reported works, the major advantage comes from the comprehensive RBFN of this work. The training effort as well as the training time required for this comprehensive RBFN with automatic learning, is found to be much lower to the above reported neural network configurations. This method is found to exhibit better species identification ability and it can be advantageously extended for identification of species of more mosquito genera.

This is the first study for proposing and using such an automatically configured RBFN for classification and identification of mosquito species based on ITS2 ribosomal DNA sequence data. Though the comprehensive RBFN model of this study is quite useful for accurate identification of mosquito species of different genera, the computational effort required for automatic configuration of RBFs and interconnection weights increases with the increase of number of genera and species. The methodology can be applied with similar success to identify the other species of *Anopheles*, *Aedes* and *Culex* genera as well as the species of other geographical varieties, if the network is trained properly by involving analogous species of such genera.

CONCLUSION

Rapid and accurate identification of mosquito species are of paramount importance for taking control measures against deadly diseases like malaria, filariasis, encephalitis, dengue and so on. Genetic identification is the conformation for any kind of biological classification. The data sequences of ITS2 region considered for mosquito species identification in this work are widely used to extract the phylogenetic relation due to their well-conserved nature in a particular species. In this work, a Radial Basis Function Network (RBFN) methodology with automatic network configuration is presented for identification of mosquito species based on the genetic pattern information content of ITS2 ribosomal DNA sequences. A number of data sequences in varying sizes of different vectors corresponding to *Anopheline, Aedes* and *Culex* are used to develop genera specific RBFN species identifiers and a comprehensive RBFN species identifier. The recall and generalization ability of the RBFN species identifiers are analyzed and further validated through bootstrap validation method. The genera specific RBFN models are found to provide accurate

identification of mosquito species of individual genera. The comprehensive RBFN species identifier is found to exhibit better species identification ability and it has the added advantage of identifying species of more mosquito genera. Since the RBFN computational method designed for species classification involving reported NCBI data base has shown effective classification efficiency, the method is expected to perform well for other data sequences if their analogues are involved in training.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

Decleared none.

APPENDIX A

RBFN Algorithm

In RBFN, the nodes in hidden layer represent radial basis functions (RBFs), which are characterized by the following:

(i) A center vector m_i in the input space, made up of cluster centers with elements m^i_j (j = 1 to n).

(ii) A distance measure to determine how far an input vector I, with elements I_j (j = 1 to n), is from the center vector m^i as defined by

$$d_i = \sqrt{\sum_{j=1}^{n} k^i{}_j (I_j - m^i{}_j)^2} \quad \textbf{(A.1)}$$

where k^i_j is the (i,j) th element of the shape matrix K defined as the inverse of the covariance matrix:

$$k^i{}_j = \frac{h^i{}_j}{\left(\sigma^i{}_j\right)^2} \quad \textbf{(A.2)}$$

where h^i_j is the correlation coefficient, and σ^i_j represents marginal standard deviation.

(iii) A Gaussian type of transfer function which transforms the Euclidian summation d_i (i = 1 to m), to give an output for each node as defined by

$$\phi(d_i) = \exp\left(\frac{-d_i^2}{\gamma^2}\right) \quad \textbf{(A.3)}$$

where γ is a real constant.

The hidden layer processes the output from the input layer using the distance measure of Eq (A.1) and the transfer function of the form given in Eq (A.3). The output of the network is a weighted sum of the outputs of ϕ (d_i) from the

hidden layer, *i.e.*

$$O = w_0 + \sum_{i=1}^{m} w_i \phi(d_i) \tag{A.4}$$

Automatic Configuration of RBFN

An efficient algorithm, namely, hierarchically self-organizing learning (HSOL) algorithm is used to automatically configure the RBFN. The HSOL algorithm automatically creates RBFs and adjusts parameter vectors of the RBFN. Further details concerning the HSOL algorithm can be referred elsewhere [34], [37], [38].

A simple way of measuring the progress of learning for a single output network is by defining the root mean square error, E_{rms} for N teaching patterns

$$E_{rms} = \sqrt{\frac{2}{N}\sum_{p=1}^{N} E_p} \tag{A.5}$$

with

$$E_p = \frac{1}{2}\left(t_p - O_p(n)\right)^2 \tag{A.6}$$

where t_p represents the desired output value defined by the pth teaching pattern, O_p represents the actual output value of the pth teaching pattern, and n represents a column vector which is a collection of all parameters associated with the output. The parameter saturation vector s is defined as

$$s(p) = \alpha \frac{\partial E_p}{\partial n} + (1-\alpha)s(p-1) \tag{A.7}$$

where α is a positive constant between 0 and 1, and p represents the p th teaching pattern presented to the network. The vector s provides the weighted average of (∂E_p / ∂n) over the horizon of learning iterations. The saturation criterion ρ is defined by the integration of the inverse of $\|s\|$ as

$$\rho(p) = \begin{cases} \rho(p-1) + b\dfrac{\sqrt{d_s}}{\|s(p)\|} & if\ p > p_0 \\ 0 & \text{otherwise} \end{cases} \tag{A.8}$$

where d_s is dimension of s, b is small positive constant representing the increment rate of ρ, and p is the delay factor defined as 1/α.

The network parameter update rules are derived from the negative gradient of the error function, Eq. (A.5).

The weights between the output and the i th RBF, w_i:

$$\Delta w_{ji} = \frac{-\partial E_p}{\partial w_{ji}} = \left(t_j - O_j\right)\phi_i \tag{A.9}$$

The j *th* element of the mean vector, m^i

$$\Delta m_j^i = \frac{-\partial E_p}{\partial m_j^i} = \sum_{j=1}^{N} k_j^i \left(x_j - m_j^i\right) \sum_{i=1}^{M} \phi_i (t - O) w_i \tag{A.10}$$

The marginal standard deviation, σ^i_j:

$$\Delta\sigma^{i}{}_{j} = \frac{-\partial E_p}{\partial \sigma^{i}{}_{j}} = \sum_{j=1}^{N} \frac{k^{i}{}_{jj}\left(x_j - m^{i}{}_{j}\right)^2}{\sigma^{i}{}_{j}} \phi_i \sum_{k=1}^{M}\left(t_k - O_k\right) w_{ki} \quad \textbf{(A.11)}$$

The correlation coefficient, h^{i}_{jk}:

$$\Delta h^{i}_{jk} = \frac{-\partial E_p}{\partial h^{i}{}_{jk}} = -\frac{1}{2} \frac{\left(x_j - m^{i}{}_{j}\right)\left(x_k - m^{i}{}_{k}\right)}{\left(\sigma^{i}{}_{j}\sigma^{i}_{k}\right)^2} \phi_i \sum_{k=1}^{M}\left(t_k - O_k\right) w_{ki} \quad \textbf{(A.12)}$$

The parameter vector of the output unit,

$n = \left\lfloor w^T, m^T, \sigma^T, h^T \right\rfloor^T$ is updated by

$$n_j^{new} = n_j^{old} + \eta \Delta n_j \quad \textbf{(A.13)}$$

where η is the positive constant called the learning rate, and $\Delta n^T = \left[\Delta w^T, \Delta m^T, \Delta\sigma^T, \Delta h^T\right]^T$.

APPENDIX B

The mean squared error (*MSE*) is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad \textbf{(B.1)}$$

where y_i and ŷ andare the target and predicted species output values and n is the number of data sequences used for prediction.

The correlation coefficient (R^2) is defined as

$$R^2 = \left\{1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}\right\} \quad \textbf{(B.2)}$$

where $\bar{y}$ is the mean value of target measurements used for prediction.

REFERENCES

[1] Wodak SJ, Janin J. Computer analysis of protein-protein interaction. J Mol Biol 1978; 124(2): 323-42. [http://dx.doi.org/10.1016/0022-2836(78)90302-9] [PMID: 712840]

[2] Dopazo J, Wang H, Carazo JM. "A new type of unsupervised growing neural network for biological sequence classification that adopts the topology of a phylogenetic tree," *Biological and Artificial Computation: Neuroscience and Technology*. Lect Notes Comput Sci 2005; 1240: 932-41. [http://dx.doi.org/10.1007/BFb0032553]

[3] Holley LH, Karplus M. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA 1989; 86(1): 152-6. [http://dx.doi.org/10.1073/pnas.86.1.152] [PMID: 2911565]

[4] Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. Bioinformatics 2002; 18(12): 1553-61. [http://dx.doi.org/10.1093/bioinformatics/18.12.1553] [PMID: 12490438]

[5] Lee K. Computational study for protein-protein docking using global optimization and empirical potentials. Int J Mol Sci 2008; 9(1): 65-77. [http://dx.doi.org/10.3390/ijms9010065] [PMID: 19325720]

[6] Kiszewski A, Mellinger A, Spielman A, Malaney P, Sachs SE, Sachs J. A global index representing the stability of malaria transmission. Am

J Trop Med Hyg 2004; 70(5): 486-98.
[PMID: 15155980]

[7] Chen B, Butlin RK, Pedro PM, Wang XZ, Harbach RE. Molecular variation, systematics and distribution of the Anopheles fluviatilis complex in southern Asia. Med Vet Entomol 2006; 20(1): 33-43.
[http://dx.doi.org/10.1111/j.1365-2915.2006.00604.x] [PMID: 16608488]

[8] Grillet ME, El Souki M, Laguna F, León JR. The periodicity of Plasmodium vivax and Plasmodium falciparum in Venezuela. Acta Trop 2014; 129: 52-60.
[http://dx.doi.org/10.1016/j.actatropica.2013.10.007] [PMID: 24149288]

[9] Hales S, Weinstein P, Woodward A. Dengue fever epidemics in the South Pacific; driven by E1 Nino south oscillation? Lancet 1996; 348(9042): 1664-5.
[http://dx.doi.org/10.1016/S0140-6736(05)65737-6] [PMID: 8962017]

[10] Lee DJ, Hicks MM, Debenham ML, *et al.* "The Culicidae of the Australian region," In: *Australian Government Publishing Service, Canberra* vol. 7, 1989.

[11] Walton C, Sharpe RG, Pritchard SJ, Thelwell NJ, Butlin RK. Molecular identification of mosquito species. Biol J Linn Soc Lond 1999; 68: 241-56. a
[http://dx.doi.org/10.1111/j.1095-8312.1999.tb01168.x]

[12] Wesson DM, Porter CH, Collins FH. Sequence and secondary structure comparisons of ITS rDNA in mosquitoes (Diptera: Culicidae). Mol Phylogenet Evol 1992; 1(4): 253-69.
[http://dx.doi.org/10.1016/1055-7903(92)90001-W] [PMID: 1364170]

[13] Marrelli MT, Floeter-Winter LM, Malafronte RS, *et al.* Amazonian malaria vector anopheline relationships interpreted from ITS2 rDNA sequences. Med Vet Entomol 2005; 19(2): 208-18.
[http://dx.doi.org/10.1111/j.0269-283X.2005.00558.x] [PMID: 15958027]

[14] Collins FH, Paskewiz SM, Finnerty V. Ribosomal RNA genes of the *Anopheles gambiae* species compex. AdvDisVector Res 1989; 6: 1-26.
[http://dx.doi.org/10.1007/978-1-4612-3292-6_1]

[15] Miller BR, Crabtree MB, Savage HM. Phylogeny of fourteen Culex mosquito species, including the *Culex pipiens* complex, inferred from the internal transcribed spacers of ribosomal DNA. Insect Mol Biol 1996; 5(2): 93-107.
[http://dx.doi.org/10.1111/j.1365-2583.1996.tb00044.x] [PMID: 8673266]

[16] Marinucci M, Romi R, Mancini P, Di Luca M, Severini C. Phylogenetic relationships of seven palearctic members of the maculipennis complex inferred from ITS2 sequence analysis. Insect Mol Biol 1999; 8(4): 469-80.
[http://dx.doi.org/10.1046/j.1365-2583.1999.00140.x] [PMID: 10634971]

[17] Marrelli MT, Sallum MAM, Marinotti O. The second internal transcribed spacer of nuclear ribosomal DNA as a tool for Latin American anopheline taxonomy -A critical review. *Mem Inst Oswaldo Cruz* 2006; 101 (8): 817-32.

[18] Sawabe K, Takagi M, Tsuda Y, Tuno N. Molecular variation and phylogeny of the Anopheles minimus complex (Diptera: Culicidae) inhabiting Southeast Asian countries, based on ribosomal DNA internal transcribed spacers, ITS1 and 2, and the 28S D3 sequences. Southeast Asian J Trop Med Public Health 2003; 34(4): 771-80.
[PMID: 15115086]

[19] Wilkerson RC, Reinert JF, Li C. Ribosomal DNA ITS2 sequences differentiate six species in the Anopheles crucians complex (Diptera: Culicidae). J Med Entomol 2004; 41(3): 392-401.
[http://dx.doi.org/10.1603/0022-2585-41.3.392] [PMID: 15185940]

[20] Bohr H, Bohr J, Brunak S, *et al.* A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. FEBS Lett 1990; 261(1): 43-6.
[http://dx.doi.org/10.1016/0014-5793(90)80632-S] [PMID: 19928342]

[21] Chen J, Chaudhari N. Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinformatics 2007; 4(4): 572-82.
[http://dx.doi.org/10.1109/tcbb.2007.1055] [PMID: 17975269]

[22] Blinder P, Baruchi I, Volman V, Levine H, Baranes D, Jacob EB. Functional topology classification of biological computing networks. Nat Comput 2005; 4: 339-61.
[http://dx.doi.org/10.1007/s11047-005-3667-6]

[23] Simpson RG, Williams R, Ellis RE, Culverhouse PF. Biological pattern recognition by neural networks. Mar Ecol Prog Ser 1992; 79: 303-8.
[http://dx.doi.org/10.3354/meps079303]

[24] Yu-Yen OU, Gromiha MM, Chen SA, Suva M. *TMBETADISC*: Discrimination of beta barrel membrane proteins using RBF networks and PSSM profile, Comput Biol Chem 2008; 32(3): 227-31.

[25] Demeler B, Zhou GW. Neural network optimization for *E. coli* promoter prediction. Nucleic Acids Res 1991; 19(7): 1593-9.
[http://dx.doi.org/10.1093/nar/19.7.1593] [PMID: 2027766]

[26] O'Neill MC. Escherichia coli promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. Nucleic Acids Res 1992; 20(13): 3471-7.
[http://dx.doi.org/10.1093/nar/20.13.3471] [PMID: 1630917]

[27] Horton PB, Kanehisa M. An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. Nucleic Acids Res 1992; 20(16): 4331-8.
[http://dx.doi.org/10.1093/nar/20.16.4331] [PMID: 1508724]

[28] Moore A. Artificial neural network trained to identify mosquitoes in flight. J Insect Behav 1991; 4(3): 391-6.
[http://dx.doi.org/10.1007/BF01048285]

[29] Banerjee AK, Kiran K, Murty USN, Venkateswarlu Ch. Classification and identification of mosquito species using artificial neural networks. Comput Biol Chem 2008; 32(6): 442-7.
[http://dx.doi.org/10.1016/j.compbiolchem.2008.07.020] [PMID: 18838305]

[30] Venkateswarlu Ch, Kiran K, Eswari JS. A hierarchical artificial neural system for genera classification and species identification in mosquitoes. Appl Artif Intell 2012; 26: 903-20.
[http://dx.doi.org/10.1080/08839514.2012.731342]

[31] Chen S, Billings SA, Covan CFN, Grant PM. Nonlinear systems identification using radial basis functions. Int J Syst Sci 1990; 21: 2513-39.
[http://dx.doi.org/10.1080/00207729008910567]

[32] Musavi MT, Ahmed W, Chan KH, Faris KB, Hummels DM. On the training of radial basis function classifiers. Neural Netw 1992; 5: 595-603.
[http://dx.doi.org/10.1016/S0893-6080(05)80038-3]

[33] Warnes MR, Glassey J, Montague GA, Kara B. Application of radial basis function and feedforward artificial neural networks to the Eschericia coli fermentation process. Neuro Comp 1998; 20: 67-82.

[34] Venkateswarlu Ch, Venkat Rao K. Dynamic recurrent radial basis function network model predictive control of unstable nonlinear processes. Chem Eng Sci 2005; 60: 6718-32.
[http://dx.doi.org/10.1016/j.ces.2005.03.070]

[35] Mayorga RV, Carrera J. A radial basis function network approach for the computation of inverse continuous time variant functions. Int J Neural Syst 2007; 17(3): 149-60.
[http://dx.doi.org/10.1142/S0129065707001020] [PMID: 17640096]

[36] Lorenz C, Ferraudo AS, Suesdek L. Artificial Neural Network applied as a methodology of mosquito species identification. Acta Tropica 2015; 152: 165-9.
[http://dx.doi.org/10.1016/j.actatropica.2015.09.011]

[37] Lee S, Kil RM. A Gaussian potential function network with hierarchically self organizing learning. Neural Netw 1991; 4: 207-24.
[http://dx.doi.org/10.1016/0893-6080(91)90005-P]

[38] Anand P, Siva Prasad BVN, Venkateswarlu Ch. Modeling and optimization of a pharmaceutical formulation system using radial basis function network. Int J Neural Syst 2009; 19(2): 127-36.
[http://dx.doi.org/10.1142/S0129065709001896] [PMID: 19496208]