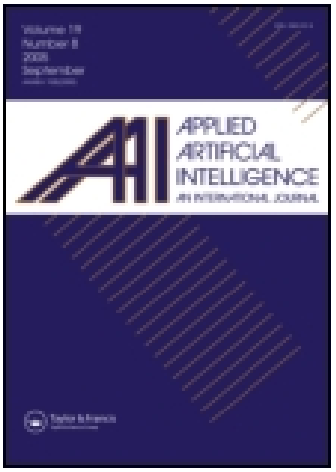

Applied Artificial Intelligence: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uaai20

A HIERARCHICAL ARTIFICIAL NEURAL SYSTEM FOR GENERA CLASSIFICATION AND SPECIES IDENTIFICATION IN MOSQUITOES

C. Venkateswarlu [a] , K. Kiran [b] & J. S. Eswari [b]

[a] Chemical Engineering Department, Padmasri Dr. B. V. Raju Institute of Technology , Narsapur , India

[b] Chemical Engineering Division, Indian Institute of Chemical Technology , Hyderabad , India

Published online: 19 Nov 2012.

To cite this article: C. Venkateswarlu , K. Kiran & J. S. Eswari (2012) A HIERARCHICAL ARTIFICIAL NEURAL SYSTEM FOR GENERA CLASSIFICATION AND SPECIES IDENTIFICATION IN MOSQUITOES, Applied Artificial Intelligence: An International Journal, 26:10, 903-920, DOI: 10.1080/08839514.2012.731342

To link to this article: http://dx.doi.org/10.1080/08839514.2012.731342

Applied Artificial Intelligence, 26:903–920, 2012

ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839514.2012.731342

Taylor & Francis
Taylor & Francis Group

A HIERARCHICAL ARTIFICIAL NEURAL SYSTEM FOR GENERA CLASSIFICATION AND SPECIES IDENTIFICATION IN MOSQUITOES

C. Venkateswarlu[1], K. Kiran[2], and J. S. Eswari[2]
[1] *Chemical Engineering Department, Padmasri Dr. B. V. Raju Institute of Technology, Narsapur, India*
[2] *Chemical Engineering Division, Indian Institute of Chemical Technology, Hyderabad, India*

□ *A hierarchical artificial neural system (HANS) is proposed for genera classification and species identification in mosquitoes, based on the internal transcribed spacer 2 (ITS2) data of ribosomal DNA sequence. The HANS is composed of two levels: the first level has a single network that serves as a genera classifier and the second level has multiple networks to perform as species identifiers. The proposed method is studied by considering a number of data sequences of various species of Aedes, Anopheles, and Culex genera. The significant feature of HANS is that when an unknown sequence is presented to the system, it first classifies the genera to which the sequence corresponds and then proceeds with the same sequence information to identify the species of the respective genera. Two more methodologies, a combined artificial neural system (CANS) and a separate artificial neural system (SANS), are also presented and compared with HANS. The results demonstrate the superior performance of HANS for genera classification and species identification in mosquitoes.*

INTRODUCTION

Bioinformatics is a rapidly developing field of research focusing on information processing of biological data through the use of computational techniques. Classification is a subject of bioinformatics where the computer is expected to analyze presented data and then make decisions based on the information content of the given data. Gene sequences in DNA are enriched with high amounts of information, but the enormity and complexity of this information makes it difficult to analyze the sequences by using traditional manual methodologies such as taxonomical classification. Accurate and efficient computational tools are needed to extract genetic-level information in the gene sequences. Mosquitoes serve as obligate intermediate hosts for numerous diseases that collectively

Address correspondence to C. Venkateswarlu, Chemical Engineering Department, Padmasri Dr. B. V. Raju Institute of Technology, Narsapur–502313, Andhra Pradesh, India. E-mail: chvenkat@iict.res.in

represent a major cause of human mortality and morbidity worldwide. There are about 34 genera with 3500 species of mosquitoes around the globe. Among them, *Anopheline, Culicine*, and *Aedine* mosquitoes are the major vectors of disease agents (Kessler and Guerin 2008; Leonard and Jan 1997). Some of the diseases for which these mosquitoes act as vectors include malaria, filaria, Japanese encephalitis, dengue, yellow fever, among others. *Anopheles* mosquitoes are spread extensively around the world (Kiszewski et al. 2004). The distribution pattern of *anophelines* is characterized by diverse geographical locations (Chen et al. 2006). Among the 3500 species of mosquitoes, approximately 430 are *Anopheles*, and of these, approximately 30–40 species act as vectors for malaria. Analyses of mosquito samples of these species have shown a high level of conservation despite being from different geographical locations with a distinguishable diversity of other species (Collins, Paskewitz, and Finnerty 1989; Fritz et al. 1994). *Aedes* is a genus of mosquito originally found in tropical and subtropical zones, but it has spread through human activity to all continents with the exception of Antarctica. The distribution of *Aedes* species is associated with climatic conditions (Hales, Weistein, and Woodward 1996). The genus of *Aedes* contains over 700 species. *Aedes* species typically are small-mosquitoes. Several of the species transmit potentially serious human diseases. *Culex* is a genus of mosquito and is important in that several species serve as vectors of diseases. The genus of *Culex* has widespread geographical distribution (Lee et al. 1989) and consists of about 767 species.

Classical methods of distinguishing species through genetic analyses (Walton et al. 1999) have reduced dependence on error-prone morphological and anatomical bases of classification (Dobzhansky 1937). In the taxonomic classification it is very difficult to classify the sibling species that bears similar morphological and anatomical features but differs in the genetic level. Therefore, for rapid identification of a mosquito vector, it is always possible to target for a conserve sequence at the genetic level, which is very much species specific and shows considerable difference even among the sibling species. The internal transcribed spacer (ITS) region is widely used in taxonomy and molecular phylogenetics (Wesson, Porter, and Collins 1992; Marrelli et al. 2005). 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* (Collins and Paskewitz 1996) and has been proved useful for differentiating between closely related species of mosquitoes (Collins and Paskewitz 1996; Crabtree, Savage, and Miller 1995; Miller, Crabtree, and Savage 1996; Marinucci et al. 1999; Hacket et al. 2000; Garros, Harbach, and Manguin 2005; Marrelli et al. 2006). This region has also been extensively targeted for species classification and phylogenetic and RNA structure-related analysis (Sawabe et al. 2003; Wilkerson, Reinert, and Li 2004; Schultz et al. 2006; Muller et al. 2007).

An efficient classification system should ideally be able to extract relevant features in the data in order to distinguish species with minimum or no misclassification. Artificial neural networks and alternatives such as radial basis function networks (Anand et al. 2009) have been used as prediction and classification tools for several applications in the field of bioinformatics (Dopazo, Huaichun, and Carazo 1997), including protein structure prediction (Bohr et al. 1990), DNA sequence analysis, and biological pattern recognition (Sabbatani 1993; Blinder et al. 2005; Simpson et al. 1992; Yu-yen et al. 2008). Because of the ease of training and the flexibility to process high amounts of information with good generalization ability, neural networks are well suited for classification and identification problems. Recently, an artificial neural network methodology has been presented for identifying a mosquito species of a single *Anopheles* genera (Amit et al. 2008). However, when species corresponding to different genera are involved, the task of distinguishing species becomes more challenging and requires an efficient methodology for classifying the genera and identifying the species. This work presents a hierarchical artificial neural system (HANS) for genera classification and species identification in mosquitoes. The HANS is composed of two levels: 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. To simplify the notation, the first-level network is denoted as genera net, and the three subnetworks are referred to as species nets. The genera net is built by using the ribosomal DNA (ITS2) sequence data corresponding to the species of the three mosquito genera considered in this study: *Aedes, Anopheles*, and *Culex*. The genera net consists of multiple input nodes to process input data sequences and two output nodes to represent the respective genera coding. Each of the second-level species nets is a multi-input and single-output network, built by using the gene sequence data of the species of each genera as its inputs and the respective species coding as the output. The schematic of the proposed HANS is shown in Figure 1. The significant feature of HANS is that when unknown sequence information is presented to the system, it first classifies the genera to which the sequence corresponds and then proceeds, with the same sequence information, to identify the species of the respective genera from the corresponding species net. Two more methodologies—a combined artificial neural system (CANS) based on a single network and a separate artificial neural system (SANS) based on genera-specific individual networks—are also presented for the analysis of gene sequences in mosquitoes. The CANS is a multi-input and single-output network, which is built by using the gene sequence data of all species of the three genera as its inputs and the respective species coding as the output. The SANS configuration involves three individual multi-input and single-output networks, each of which is built by using the gene sequence information

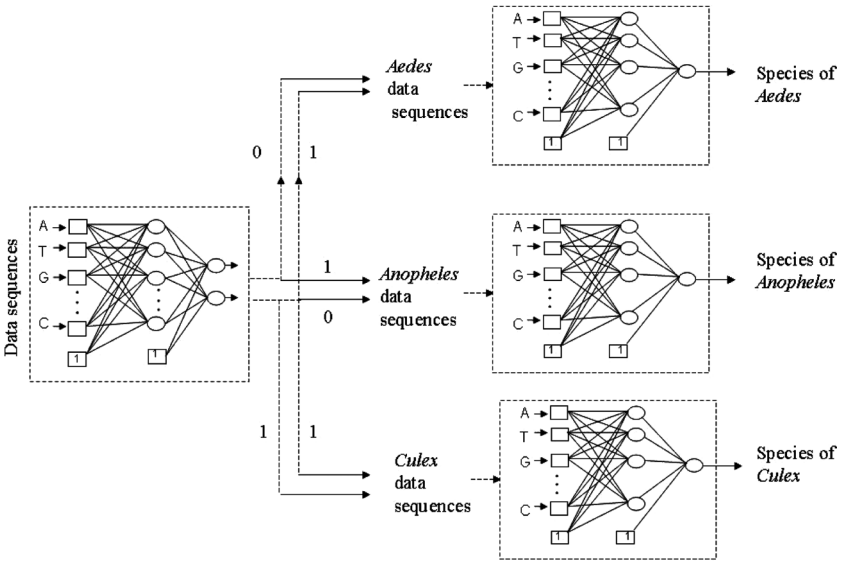

FIGURE 1 Structure of HANS. This structure is composed of two levels: the first level has a single network to serve as a genera classifier and the second level has multiple networks to perform as species identifiers.

corresponding to the species of each genera as its inputs and the respective species coding as the output. The structure of CANS is similar to that used for mosquito species identification of a single genera (Amit 2008; Nature India). Each of the subnets in SANS is genera specific and is assigned with the structure similar to CANS. All these methodologies are presented in detail and their performances are evaluated in terms of their generalization ability for better genera classification and rapid species identification. The programs corresponding to these methodologies are executed in C using the code written by the authors.

ALGORITHM

Artificial Neural Networks

Artificial neural networks (ANNs) are computer systems developed to mimic the operations of the human brain by mathematically modeling its neurophysiological structure. ANNs consist of a large number of computational units connected in a massively parallel structure. These computational units are called *neurons* and represent the nerve cells in the brain, and the strengths of the interconnections are represented by *weights*, in which the learned information is stored. Collectively, the interconnections between the layers of the network and the transfer functions of the

processing units can form distributed representations of the relationships between input and output data. This unique arrangement can acquire some of the neurological processing ability of the biological brain, such as learning and drawing conclusions from experience. The widely used ANN paradigm is a multilayered, feed-forward network (MFFN) with a multilayered perceptron (MLP), mostly comprising three sequentially arranged layers of processing units (Jones 1987; Girosi and Poggio 1990). The MFFN provides a mapping between an input (x) and an output (y) through a nonlinear function f, in other words, $y=f(x)$. The three-layered MFFN has input, hidden, and output layers, each layer comprising its own nodes. All the nodes in the input layer are connected using weighted links to the hidden-layer nodes; similar links exist between the hidden- and output-layer nodes. Usually, the input and hidden layers also contain a bias node possessing a constant output of 1. The nodes in the input layer do not perform any numerical processing; all numerical processing is done by the hidden- and output-layer nodes, and they are termed *active* nodes.

Training Algorithm

The problem of neural network training is to obtain a set of weights such that the prediction error defined by the difference between the network's predicted outputs and the desired outputs is minimized. The iterative training causes the network to recognize patterns in the data and creates an internal model that provides predictions for the new input condition. The input to the network consists of n-dimensional vector x_p and a unit bias. Each input is multiplied by a weight w_{ij}, and the products are summed to obtain the activation state S_{pj} as shown:

$$S_{pj} = \sum_{i=1}^{N} w_{ij} x_{pi} + w_{N+1,j}. \tag{1}$$

The output of the hidden-layer neuron O_{pj} for sigmoid function is calculated as

$$O_{pj} = f(S_{pj}) = \frac{1}{1 + e^{-S_{pj}}}, \tag{2}$$

where f represents the differentiable and nondecreasing function. The output layer of a single hidden-layer network performs the same calculations as previously described, except that the input vector x_p is replaced by the hidden-layer output O_p and the corresponding weights w_{jk}:

$$S_{pk} = \sum_{i=1}^{M} w_{jk} O_{pi} + w_{M+1,k}. \tag{3}$$

$$O_{pk} = y_{pk} = \frac{1}{1 + e^{-S_{pk}}}. \tag{4}$$

Similar calculations can be extended to networks containing more than one hidden layer.

A simple way to measure the progress of learning is by defining the sum of squared error E_p for p learning patterns. The set of training examples consists of p input-output vector pairs (x_p, d_p). Weights are initially randomized. Thereafter, weights are adjusted in order to minimize the objective function $E(w)$, defined as the mean squared error between the prediction outputs y_{pk} and the target outputs d_{pk} for all the input patterns:

$$E(w) = \sum_{p=1}^{p} E_p, \tag{5}$$

where E_p is the sum of the squared error with each training example,

$$E_P = \sum_{K=1}^{M} (d_{pk} - y_{pk})^2. \tag{6}$$

The task of E_p minimization is accomplished by training the network using a gradient descent technique such as the generalized delta rule (Jones and Hoskins 1987; Girosi and Poggio 1990). According to this rule, the error function δ_{pk} between the hidden-layer neurons and the output-layer neuron k is computed:

$$\delta_{pk} = (d_{pk} - y_{pk}) f'(S_{pk}). \tag{7}$$

The error function δ_{pj} from input neuron to hidden neuron can be calculated as

$$\delta_{pj} = f'(S_{pj}) \sum_{k=1}^{M} \delta_{pk} w_{jk}. \tag{8}$$

The weight change Δw from output to hidden layer after nth data presentation is given by

$$\Delta w_{jk}(n) = \eta \delta_{pk} O_{pk} + \alpha \Delta w_{jk}(n-1), \tag{9}$$

where η is the learning rate and α is the momentum factor. The updated weights are given by

$$w_{jk}(n) = w_{jk}(n-1) + \Delta w_{jk}(n). \tag{10}$$

The weight changes from hidden to input layer can be calculated in the same way. After the weights are updated, a new training example is randomly

selected, and the procedure is repeated until satisfactory reduction of the objective function is achieved.

Information Processing

Network training is an iterative procedure that begins with initializing the weight matrix randomly. Network learning involves two types of passes: a forward pass and a reverse pass. In the forward pass, an input pattern from the training data set is applied to the input nodes, the weighted sum of the inputs to the active node is calculated and is then transformed into output using a nonlinear activation function such as the sigmoid function. The outputs of the hidden nodes computed in this manner form the inputs to the output-layer nodes whose outputs are evaluated in a similar fashion. In the reverse pass, the pattern-specific squared error defined in terms of target and prediction outputs is used for updating the network weights. The weight updating procedure when repeated for all the patterns in the training set completes one iteration. For a given ANN-based modeling problem, the number of nodes in the network input layer and output layer are dictated by the input-output dimensionality of the pattern being modeled. However, the number of hidden nodes is an adjustable structural parameter. If the network architecture contains more hidden units than necessary, it leads to an oversized network. To avoid over-fitting of the network, the network simulations are to be conducted by systematically varying the number of hidden units. These simulations provide optimal network architecture with the smallest error magnitude for the test data.

METHODS

ANN Methodologies for Genera Classification and Species Identification

ANNs with backpropagation represent the most popular learning paradigm and have been successfully used to perform a variety of input-output mapping tasks for recognition, generalization, and classification (Dayhoff 1990). As a technique for computational analysis, neural network technology is very well suited for the analysis of molecular sequencing data. In recent years, backpropagation neural networks have been used to predict secondary and tertiary protein structures (Holley, and Karplus 1989; Kneller, Cohen, and Langridge 1990) to detect DNA binding sites (O'Neill 1991), to predict bacterial promoter sequences (Demeler and Zhou 1991) and as a protein classification system (Wu et al. 1992). The principal aim of this work is to present a novel neural network methodology for accurate classification of mosquito genera and rapid identification of the mosquito species based on

the genetic information present in ITS2 gene sequences of the species of different genera.

Data Collection

The ribosomal DNA sequence (ITS2) data of 34 species of the three genera *Anopheles, Aedes*, and *Culex* is collected in a fast format from the National Center for Biotechnology Information (NCBI) nucleotide database (Genbank, at www.ncbi.nlm.nih.gov/) and is used for computational experiment. The length of the sequences varies in the range of 200 to 500. Each sequence collected from the NCBI nucleotide database has its own accession number. For example, the ITS2 part of the representative sample sequence of *Anopheles saporoi* belonging to the *Anopheles* genera has the NCBI accession number AY425338.1, the sample sequence of *Aedes albiradius* belonging to the *Aedes* genera has the NCBI accession number FM211137.1, and the sample sequence of *Culex pipens* belonging to the *Culex* genera has the NCBI accession number AM084683.1. Additional information concerning the genera and the species, together with their representative coding used for the development of ANN methodologies, is given in the following sections.

Data Encoding

Data encoding plays a crucial role in enhancing the network performance. Neural networks are able to process many different forms of data as long as they are presented to the network in an acceptable format. The encoding methodology used should be representative of the data such that the neural network can make clear distinctions among the different classes of inputs. ITS2 ribosomal DNA sequences of mosquito species are made up of four bases, A, T, G, and C. These four bases should be represented in the form of a numerical vector in order to train and use a neural network for genera classification and sequence classification. The numerical values to encode the input data sequences and to represent the output genera and species are chosen in order to satisfy the training, prediction, and classification needs of the neural network methodologies presented in the following sections.

Hierarchical Artificial Neural System (HANS)

The structure of HANS is shown in Figure 1. The HANS consists of a genera net and three species nets. The bases Adenine-A, Thyamine-T, Guanine-G, and Cytosine-C of ribosomal DNA sequences represent the inputs to the genera net with their corresponding genera names as the

network outputs. The bases A, T, G, and C of the network inputs are coded with binary values as A = {0 0}, T = {0 1}, G = {1 0}, and 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 plus a bias node. The output coding used to represent the genera names is a binary coding with {0 1} for *Aedes*, {1 0} for *Anopheles*, and {1 1} for *culex*. The input coding assigned for the gene sequences of each of the species nets is the same as the genera net. The output coding used to represent the species names in each of the species nets is a real coding. The output coding assigned for the genera net and the three species nets in HANS is given in Table 1 together with the respective genera and species names. The coding in Table 1 is selected based on several trials of different input and output coding combinations. A total of 34 species of the three genera—with 14 species of *Anopheles*, 10 species of *Aedes*, and 10 species of *Culex*—are considered for building the HANS. Altogether, six sequences of each species (i.e., 204 sequences) of variable lengths are used for training and testing the HANS. Out of these sequences, three random sequences of each species (i.e., 102 sequences) are used to train the HANS, and an equivalent number of random sequences are considered to evaluate the performance of the trained model. In order to accommodate the input data sequences of varying lengths, the number of input nodes to the genera net in HANS is specified with the binary coding equivalent to that of a larger sequence. With this network input node configuration, it is convenient to appropriately fill the additional nodes remaining from shorter sequences, thus to accommodate a sequence of any length. The network training is performed by using a backpropagation delta-rule algorithm. The network interconnection weights are initialized by assigning random numbers in the range of 0.1 to −0.1. The genera net training involves the sequential treatment of the data sets corresponding to the input sequences of the three genera to map with the output coding specified for the genera. The iterative convergence of the minimum objective causes the network to learn and modify the values of interconnection weights between the nodes of the layers. This minimization criterion also enables the selection of the number of hidden nodes and the number of iterations required for training the genera net. The procedure used to train the species nets is similar to that of the genera net. The input data sequences for the species of each genera in binary coding and the corresponding output species names in real coding are used sequentially to train each of the species nets. Input and output mapping comparison of target and actual values continue until all mapping sequences of the training species are learned within an acceptable overall error. The number of hidden units, the number of iterations, the learning rate α, and the momentum factor η in genera net and the three species nets are suitably selected. During the

TABLE 1 Output Coding for HANS and CANS

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

classification and generalization phase, the trained HANS model itself operates in a feed-forward manner.

Combined Artificial Neural System (CANS)

The CANS is a multi-input and single-output network, in which all the data sequences corresponding to the species of the three genera are treated in a single network. As in HANS, the bases A, T, G, and C of ribosomal DNA sequences form the network inputs, and their corresponding species names, represent the network outputs. The bases in input sequences are assigned with binary coding as A = {0 0}, T = {0 1}, G = {1 0}, C = {1 1},

and the outputs are specified by real coding. The output coding for all the species of the three genera considered in this study is shown in Table 1. As in HANS, 102 random sequences from the 34 species of the three genera are used to train the CANS, and an equivalent number of random sequences of the species of the three genera is used to evaluate the performance of the trained model. The number of input nodes used to accommodate the sequences of varying lengths in CANS is the same as in HANS. The network training is performed using a backpropagation delta-rule algorithm. The initialization of the network interconnection weights in CANS is the same as in HANS. 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 in order to model 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. The weight adjustment enforced by the learning rule propagates exactly backward from the output layer through the hidden layer toward the input layer. During the classification phase, the trained CANS itself operates in a feed-forward manner.

Separate Artificial Neural System (SANS)

This configuration consists of three separate neural networks, each representing one genera. Each of the networks in SANS is a multi-input, single-output network, which is built by using the gene sequences corresponding to the species of each genera as its inputs and the respective species names as the output. These three individual networks in SANS are the same as the three species nets in HANS. However, in SANS, these networks operate independently, whereas in HANS they are associated with the first-level genera net, and these serve as second-level species classifiers. The genera-specific individual models in SANS operate in a feed-forward manner during the recall and generalization phases.

RESULTS AND DISCUSSION

The development of HANS, CANS, and SANS is carried out according to the procedure detailed in the above sections. The genera net in HANS is built by sequentially processing the input gene sequences of the species of the three genera and mapping the model predictions with the respective genera-specific output data. Each of the species nets in HANS is built by sequentially processing the input data sequences corresponding to the species of each genera and mapping the predictions with the data

corresponding to the output species of the respective genera. In order to accommodate the input data sequences of the species of varying lengths, the network input nodes are assigned with a size equivalent to that of a maximum sequence length and the shorter sequences are accommodated by repeatedly filling the additional input nodes with the binary coding of the last four bases in the sequence. Both the genera net and species nets in HANS are trained by using the backpropagation delta-rule algorithm described in the earlier section. The CANS is trained by sequentially treating the input data sequences corresponding to the species of the three genera along with the respective species output coding. Each of the genera-specific networks in SANS is configured by using the data sequences corresponding to the species of the respective genera. A total 102 random sequences of the three genera are used to train the networks involved in HANS, CANS, and SANS. The number of hidden units, the number of iterations, the learning rate α, and the momentum factor η for each of these configurations are selected such that the trained models with these parameters satisfy the recall and generalization ability. The training is thus performed in order to establish convergence in the objective functions of the respective network configurations. The parameters chosen for the three modeling configurations and their training performances are given in Table 2. Because the three individual networks representing *Aedes, Anopheles,* and *Culex* in SANS are same as the three species nets in HANS, the training parameters of these networks are the same as the species nets in HANS. Each of the networks in HANS and CANS with four hidden nodes is found to converge with minimum training error. As shown in Table 2, the number of iterations and the training time required for CANS is found to be more than that of the genera net and the species nets in HANS. For each network configuration, the iterative convergence of the normalized error with respect to the number of iterations is shown in Figure 2. The results in Figure 2 indicate the faster convergence of the species nets and the genera net in HANS than that in CANS. The programs corresponding to these methodologies are implemented in C language on a personal computer (Pentium 4 CPU, 3.2 GHz, 512 MB of RAM).

TABLE 2 Training Parameters and Training Performance

	HANS				
Training parameters	Genera net	Species net1	Species net2	Species net3	CANS
α	0.820	0.782	0.820	0.782	0.820
η	0.00179	0.00179	0.00179	0.00179	0.0059
Iterations	55000	20,000	20,000	20,000	1,50,000
Hidden nodes	4	4	4	4	4
Minimum training error	0.072816	0.006544	0.008419	0.006620	0.002428
Training time	10.7 min	0.62 min	0.65 min	0.98 min	28.3 min

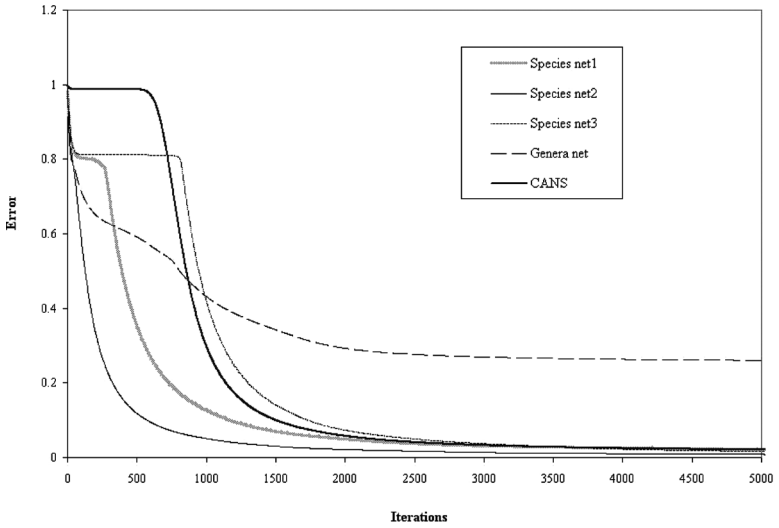

FIGURE 2 Iterative error convergence. The comparison of error convergence with respect to the number of iterations for the genera net and species nets in HANS and the error convergence of CANS. The number of iterations is limited to 5000, for convenience.

The well-trained and learned models are then assessed for their recall and generalization abilities. The recall abilities of HANS, CANS, and SANS are evaluated by using the same input sequences as those employed for training. All these modeling configurations have exhibited 100% recall ability. However, the predictive performances of these models rely on their generalization ability. The genera classification and species identification efficiency of these models are then tested by using the species sequences that are not involved in training. A total of 102 random sequences of the three genera are used to test HANS, CANS, and SANS configurations. When an untrained DNA sequence of a species is presented as input, the HANS could be able to perfectly classify the genera and exactly identify the species corresponding to that genera. The genera net in HANS first identifies the genera to which the data sequence belongs and then sends the same input information to the respective species net to identify the species in that genera. Thus, the first stage of HANS enables classification of the genera, and the second stage identifies the species of that genera. The species identification results of HANS for the untrained sample sequences are compared with their corresponding target values as depicted in Figure 3. The close concurrence between the prediction results and the target values in Figure 3 indicates better genera classification and species identification ability of HANS. Figure 4 shows the prediction results of CANS that are evaluated using the data sequences that are not involved in training. These results show that a few of the species predicted by CANS are not very close to their corresponding

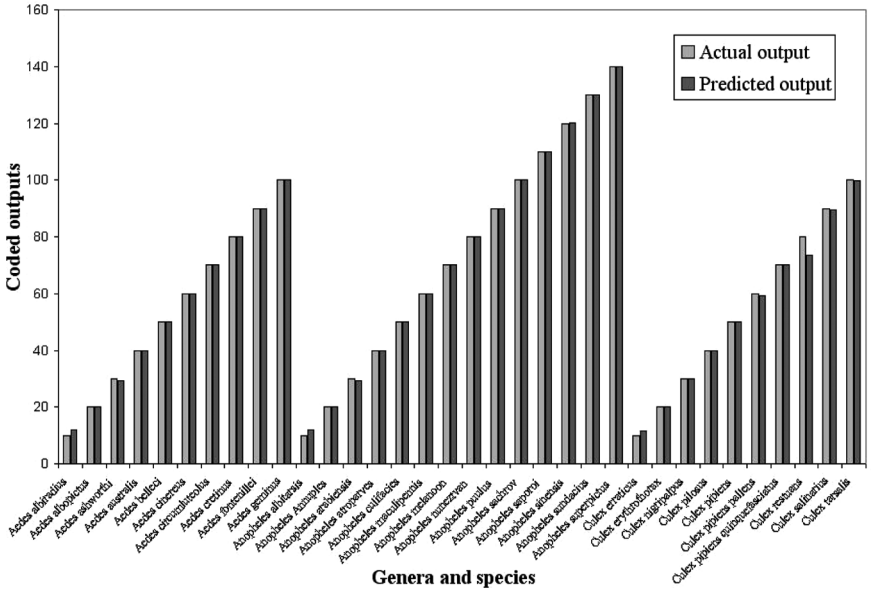

FIGURE 3 Comparison of actual and predicted outputs of HANS. The dark columns in the figure refer the predictions of the species nets for the same input sequences of the species of the genera identified by the genera net.

target's values. Each individual network in SANS corresponds to a specific genera and is trained using the input data sequences corresponding to the respective genera. The three subnets in SANS provide better predictive

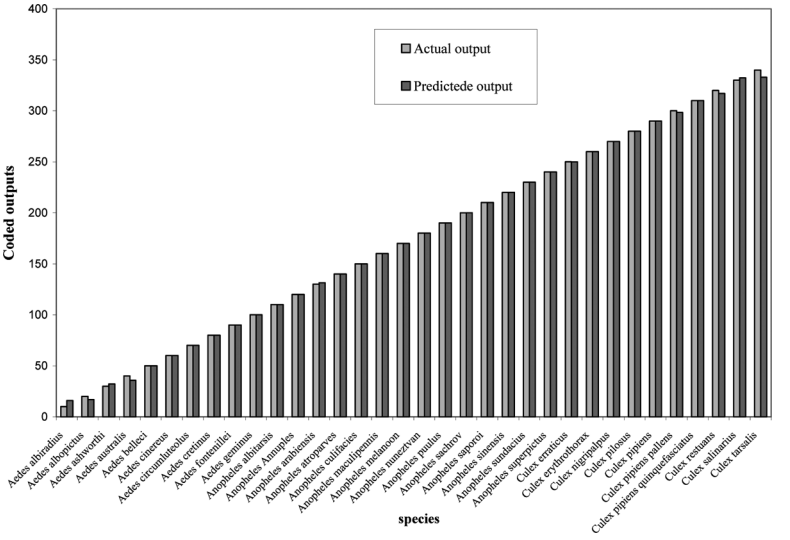

FIGURE 4 Comparison of actual and predicted outputs of CANS. The dark columns in the figure refer to the species predictions based on the input data sequences of different genera.

performance when they are used with the input data sequences of the respective genera. However, when an unknown, untrained data sequence is used as input to these three subnets, the resulting outputs with their distinct magnitudes are as shown in Figure 5. Although one of these outputs is specific to the genera for which the input sequence belongs, there is a possibility of misidentification of the species because of the presence of the other outputs in significant magnitudes. The reason is that the subnet trained based on the species data of a specific genera may not accommodate the species information concerning the other genera.

The generalization ability of HANS, CANS, and SANS is also evaluated by using quantitative performance measures, namely, mean squared error (MSE) and correlation coefficient (R^2), which are define by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{11}$$

$$R^2 = \left\{1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}\right\}, \tag{12}$$

where y_i and $\hat{y}_i$ are the target and predicted outputs, $\bar{y}$ is the mean value of target data and n represents the number of data sequences. These quantification measures for each of the modeling configurations are evaluated based on the untrained data sequences. The normalized data of model predictions and the

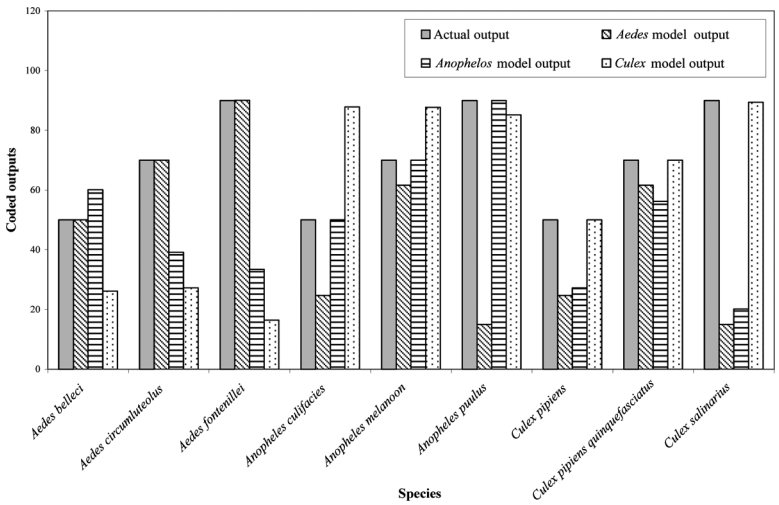

FIGURE 5 Comparison of actual and predicted outputs of SANS. The columns in the figure refer the output predictions when unknown input sequences are fed to the three genera-specific networks.

normalized target data are used to compute these measures. The MSE and R^2 values evaluated based on 102 untrained data sequences for HANS are 0.000121 and 0.997199, respectively, and those computed for CANS are 0.000173 and 0.98783, respectively. These results indicate the better predictive performance of HANS over CANS. HANS require less than half of the training time as that of CANS, and it has faster prediction ability. The CANS also requires more training effort because of the treatment of all the data sequences corresponding to the three genera in a single network. Though the subnets in SANS provide better quantitative performance when they are used with genera-specific data, they are not well suited for the treatment of unknown, untrained data sequences of other genera, as shown in Figure 5. The significant feature of HANS is that when unknown sequence information is presented to the system, it first classifies the genera to which the sequence corresponds and then proceeds with the same sequence information to identify the species of the respective genera from the corresponding species subnet.

CONCLUSIONS

Genera classification and species identification is of paramount importance for taking control measures against deadly diseases spread by mosquitoes. Methodologies based on genetic-level information of conserved data sequences can be effective for distinguishing and identifying species of biological systems. The data sequences of the ITS2 region are widely used to extract the phylogenetic relation because of their well-conserved nature in a particular species. The principal aim of this work is to present an efficient methodology for genera classification and species identification in mosquitoes based on the genetic information of the ITS2 sequences. A hierarchical artificial neural system (HANS) in two levels is presented in which the first level has a single network to serves as a genera classifier, and the second level has multiple networks to perform as species identifiers. The development and performance evaluation of the proposed method is studied by considering a number of data sequences of various species associated with different genera such as *Aedes, Anopheles*, and *Culex*. Two more methodologies—a combined artificial neural system (CANS) based on a single network, and a separate artificial neural system (SANS) based on genera-specific individual networks—are also presented and evaluated to compare them with the HANS. The artificial neural methodology, HANS, presented in this work provides accurate classification of genera and rapid identification of species in mosquitoes based on ITS2 ribosomal DNA sequences. This methodology can be effectively used for classification and identification problems related to any kind of biological system based on genetic-level information.

REFERENCES

Das, B. 2008. AI to identify mosquitoes. *Nature India* doi: 10.1038/nindia.2008.334.

Amit, B., K. Kiran, U. S. N. Murthy, and Ch. Venkateshwarlu. 2008. Classification and identification of mosquitoe species using artificial neural networks. *Journal of Computational Biology and Chemistry* 32:442–447.

Anand, P., B. V. N. Siva Prasad, and Ch. Venkateshwarlu. 2009. Modeling and optimization of a pharmaceutical formulation system using radial basis function network. *International Journal of Neural Systems* 19:127–136.

Blinder, P., L. Baruchi, V. Volman, H. Levine, D. Baranes, and E. B. Jacob. 2005. Functional topology classification of biological computing networks. *Natural Computing* 4:339–361.

Bohr, H., J. Bohr, S. Beunak, R. M. J. Cotterill, H. Fredholm, B. Lautrup, and S. B. Peterson. 1990. A novel approach to prediction of the 3 dimensional structures of protein backbones by neural network. *FEBS Letters* 261:43–46.

Chen, B., R. K. Butlin, P. M. Pedro1, X. Z. Wang, and R. E. Harbach. 2006. Molecular variation, systematics and distribution of the *Anopheles fluviatilis* complex in southern Asia. *Medical and Veterinary Entomology* 20:33–43.

Collins, F. H., and S. M. Paskewitz. 1996. A review of the use of ribosomal DNA (rDNA) to differentiate among cryptic *Anopheles* species. *Insect Molecular Biology* 5:1–9.

Collins, F. H., S. M. Paskewitz, and V. Finnerty. 1989. Ribosomal RNA genes of the *Anopheles gambiae* species complex. *Advances in Disease Vector Research* 6:1–28.

Crabtree, M. B., H. M. Savage, and B. R. Miller. 1995. Development of a species diagnostic polymerase chain reaction assay for the identification of *Culex* vectors of St. Louis encephalitis virus based on sequence variation in ribosomal DNA spacers. *American Journal of Tropical Medicine and Hygiene* 53:105–109.

Dayhoff, J. E. 1990. *Neural network architectures.* New York: Van Nostrand Reinhold.

Demeler, B., and G. Zhou. 1991. Neural network optimization for *E. Coli* promoter prediction. *Nucleic Acids Research* 19:1593–1599.

Dobzhansky, Th. 1937. Genetic nature of species differences. *The American Naturalist* 71:404–420.

Dopazo, J., W. Huaichun, and J. M. Carazo. 1997. A new type of unsupervised growing neural network for biological sequence classification that adopts the topology of a phylogenetic tree. *Lecture Notes in Computer Science* 1240:932–941.

Fritz, G. N., J. Conn, A. F. Cockburn, and J. A. Seawright. 1994. Sequence analysis of the ribosomal DNA internal transcribed spacer 2 from populations of *Anopheles nuneztovari* (Diptera: Culicidae). *Molecular Biology and Evolution* 11:406–416.

Garros, C., R. E. Harbach, and S. Manguin. 2005. Morphological assessment and molecular phylogenetics of the Funestus and Minimus groups of *Anopheles* (*Cellia*). *Journal of Medical Entomology* 42:522–536.

Girosi, F., and T. Poggio. 1990. Networks and the best approximation property. *Biological Cybernetics* 63:169–179.

Hales, S., P. Weistein, and A. Woodward. 1996. Dengue fever epidemics in the South Pacific; driven by E1 Nino south oscillation? *Lancet* 348:1664–1665.

Hacket, B. J., J. Gimnig, W. Guelbeogo, C. Constantini, L. L. Koekemoer, M. Coetzee, F. H. Collins, and N. J. Besansky. 2000. Ribosomal DNA internal transcribed spacer (ITS2) sequences differentiate *Anopheles funestus* and *An. Rivulorum*, and uncover a cryptic taxon. *Insect Molecular Biology* 9:369–374.

Holley, L. H., and M. Karplus. 1989. Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the USA* 86:152–156.

Jones, W. P., and J. Hoskins. 1987. Back Propagation. *Byte* 0:155–162.

Kessler, S., and P. M. Guerin. 2008. Responses of *Anopheles gambiae, Anopheles stephensi, Aedes aegypti*, and *Culex pipens* mosquitoes (*Dipteria: Culicidae*) to cool and humid refugium conditions. *Journal of Vector Ecology* 1:145–149.

Kiszewski, A., A. Mellinger, A. Spielman, P. Malaney, S. E. Sachs, and J. Sachs. 2004. A global index representing the stability of malaria transmission. *American Journal of Tropical Medicine and Hygiene* 5: 486–498.

Kneller, D. G., F. E. Cohen, and R. Langridge. 1990. Improvement in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology* 214:171–182.

Lee, D. J., M. M. Hicks, M. L. Debenham, M. Griffins, E. N. Marks, J. H. Bryan, and R. C. Russel. 1989. The *Culicidae* of the Australian region. Canberra, AU: Australian Government Publishing Service.

Leonard, E. M. M., and E. C. Jan. 1997. Systematics of mosquito disease vectors (*Diptera, Culicidae*): Impact of molecular biology and cladistic analysis. *Annual Review of Entomology* 42:351–369.

Marrelli, M. T., M. A. M. Sallum, and O. Marinotti. 2006. The second internal transcribed spacer of nuclear ribosomal DNA as a tool for Latin American anopheline taxonomy -A critical review. *Memorias do Instituto Oswaldo Cruz* 8:817–832.

Marrelli, M. T., L. M. Floeter-Winter, R. S. Malafronte, W. P. Tadei, L. R. Lourenço-de-oliveira, C. Flores-Mendoza, and O. Marinotti. 2005. Amazonian malaria vector *anopheline* relationships interpreted from ITS2 rDNA sequences. *Medical and Veterinary Entomology* 2:208–218.

Marinucci, M., R. Romi, P. Mancini, M. DiLuca, and C. Severini. 1999. Phylogenetic relationships of seven palearctic members of the *maculipennis* complex inferred from ITS2 sequence analysis. *Insect Molecular Biology* 4:469–480.

Miller, B. R., M. B. Crabtree, and H. M. Savage. 1996. Phylogeny of fourteen *Culex* mosquito species, including the *Culex pipens* complex, inferred from the interal transcribed spacers of ribosomal DNA. *Insect. Molecular Biology* 2 (1996): 93–107.

Muller, T., N. Philippi, T. Dandekar, J. Schultz, and M. Wolf. 2007. Distinguishing species. *RNA* 13: 1469–1472.

O'Neill, M. C. 1991. Training back propagation neural networks to define and detect DNA binding sites. *Nucleic Acids Research* 19 (1991): 313–318.

Sabbatini, R. M. E. 1993. Neural networks for classification and pattern recognition of biological signals. *Proceedings of the 15th annual international conference of the IEEE.* 265–266. San Diego, CA, USA.

Sawabe, K., M. Takagi, Y. Tsuda, and N. Tuno. 2003. Molecular variation and phylogeny of the *Anopheles minimus* complex (*Diptera:Culicidae*) inhibiting southeast asian countries, based on ribosomal DNA internal transcribed spacers, ITS1 and ITS2, and the 28S D_3 sequences. *Southeast Asian Journal of Tropical Medicine and Public Health* 4: 771–780.

Schultz, J., T. Muller, M. Achtziger, P. N. Seibel, T. Dandekar, and W. Matthias. 2006. The internal transcribed spacer 2 database-a web sever for (not only) low level phylogenetic analyses. *Nucleic Acids Research* 34:704–707.

Simpson, R. G., R. Williams, R. E. Ellis, and P. F. Culverhouse. 1992. Biological pattern recognition by neural networks. *Marine Ecology Progress Series* 79:303–308.

Walton, C., R. G. Sharpe, S. J. Pritchard, N. J. Thelwell, and R. K. Butlin. 1999. Molecular identification of mosquito species. *Biological Journal of the Linnean Society* 68:241–256.

Wesson, D. M., C. H. Porter, and F. H. Collins. 1992. Sequence and secondary structure comparisons of ITS rDNA in mosquitoes (*Diptera: Culicidae*). *Molecular Phylogeny and Evolution* 1:253–269.

Wilkerson, R. C., J. F. Reinert, and C. Li. 2004. Ribosomal DNA ITS2 sequences differentiate six species in the *Anopheles crucians* complex (*Diptera: Culicidae*). *Journal of Medical Entomology* 41:392–401.

Wu, C. G., J. Whitson, A. E. McLarty, and T. C. Chang. 1992. Protein classification artificial neural system. *Protein Science* 5:667–677.

Yu-Yen, O. U., M. M. Gromiha, S.-A. Chen, and M. Suva. 2008. TMBETADISC: Discrimination of beta barrel membrane proteins using RBF networks and PSSM profiles. *Journal of Computational Biology and Chemistry* 32 (1): 227–231.