Computational Biology and Chemistry 32 (2008) 442–447

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Brief Communication

Classification and identification of mosquito species using artificial neural networks☆

Amit Kumar Banerjee[a], K. Kiran[b], U.S.N. Murty[a], Ch. Venkateswarlu[b,*]

[a] Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Hyderabad 500007, Andhra Pradesh, India
[b] Chemical Engineering Sciences Division, Indian Institute of Chemical Technology, Hyderabad 500007, Andhra Pradesh, India

ARTICLE INFO

Article history:
Received 7 August 2007
Received in revised form 10 July 2008
Accepted 10 July 2008

Keywords:
Artificial neural network
Anopheles
Internal transcribed spacer 2
Mosquitoes
Malaria

ABSTRACT

An artificial neural network method is presented for classification and identification of *Anopheles* mosquito species based on the internal transcribed spacer2 (ITS2) data of ribosomal DNA string. The method is implemented in two different multi-layered feed-forward neural network model forms, namely, multi-input single-output neural network (MISONN) and multi-input multi-output neural network (MIMONN). A number of data sequences in varying sizes of different *Anopheline* malarial vectors and their corresponding species coding are employed to develop the neural network models. The classification efficiency of the network models for untrained data sequences is evaluated in terms of quantitative performance criteria. The results demonstrate the efficiency of the neural network models to extract the genetic information in ITS2 sequences and to adapt to new data. The method of MISONN is found to exhibit superior performance over MIMONN in distinguishing and identification of the mosquito vectors.

1. Introduction

Gene sequences are enriched with high amount of information, but the enormity and complexity of this information makes difficult to analyze it by using traditional manual methodologies. Efficient and intelligent computational tools are needed to extract biological relevant features in the gene sequences. In this context, computational tools such as neural networks, fuzzy logic and genetic algorithms are widely employed. The main characteristic of all these systems is their ability to learn and adapt according to the presented data.

Mosquito is the worst enemy of mankind since dawn of time and act as vector for several diseases, namely, malaria, filaria, Japanese encephalitis, dengue and so on (Leonard and Jan, 1997). Among these diseases, burden of malaria is quite considerable and result in heavy toll of life. *Anopheles* mosquito species act as a vector for malaria which is extensively spread throughout the globe (Kiszewski et al., 2004). The distribution pattern of *Anophelines* is characterized by diverse geographical locations (Chen et al., 2006). Analyses of mosquito samples have shown a high level of conservation despite of being from different geographical locations with a distinguishable diversity with other species (Fritz et al., 1994).

Classical methods of distinguishing species through genetic analyses (Walton et al., 1999a) have reduced dependence on error prone morphological and anatomical basis 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 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 et al., 1992; Marrelli et al., 2006). The internal transcribed spacer2 (ITS2) region located between the 5.8S and 28S gene is highly conserved and species specific. This region is widely used for DNA sequencing in mosquito genera of *Anopheles*, *Culex* and *Aedes* (Collins et al., 1989) and is extensively targeted for species classification, phylogenetic and RNA structure related analysis (Marinucci et al., 1999; Marrelli et al., 2005; Wilkerson et al., 2004; Banerjee et al., 2007).

Neural networks have numerous applications in the field of bioinformatics (Dopazo et al., 1997), which include protein structure prediction (Bohr et al., 1990), DNA sequence analysis and biological pattern recognition (Sabbatini, 1993; Blinder et al., 2005; Simpson et al., 1992). Due to ease of training and flexibility to process higher amount of information with good generalization ability,

☆ IICT Communication No. 070604.

* Corresponding author at: Biology Division, Indian Institute of Chemical Technology, Hyderabad 500007, Andhra Pradesh, India. Tel.: +91 40 27193121; fax: +91 40 27193626.
E-mail address: chvenkat@iict.res.in (Ch. Venkateswarlu).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.07.020

Nomenclature

d_{pk}	target value for pth teaching pattern
E_p	sum of squared error
M	number of hidden units
N	number of input units
O_{pj}	output from hidden layer neuron; output from output layer neuron
S_{pj}	activation state for hidden node
S_{pk}	activation state for output node
w_{ij}	weights between input and hidden layers
w_{jk}	weights between hidden and output layers
y_{pk}	node output for pth teaching pattern

Greek symbols

α	momentum factor
δ_{pk}	error function between the hidden and output layer neurons
δ_{pj}	error function between the input and hidden layer neurons
η	learning rate

neural networks are well suited for classification problems. This work presents an artificial neural system for classification and identification of *Anopheles* mosquito species based on the information content of ribosomal DNA sequences. Two different multi-layered feed-forward neural networks are configured for the analysis of ITS2 data sequences. The first one is a multi-input single-output neural network (MISONN) configuration, where the input nodes of the network are assigned with sequence information of various species while the output node is specified by the species coding. The second one is a multi-input multi-output neural network (MIMONN) configuration, which contains the input information as in MISONN but it has multiple output nodes for all species with their corresponding species codes. Both the network models are presented in detail and their performance is evaluated in terms of their generalization ability for better classification and rapid identification of *Anopheles* genera. The programs corresponding to these network models are executed in C using the code written by the authors. To the knowledge of the authors, the artificial neural methodology presented in this work for identification of mosquito species based on ITS2 ribosomal DNA sequences is not reported earlier.

2. Artificial Neural Networks

Artificial neural networks (ANNs) are computer systems developed to mimic the operations of the human brain by mathematically modeling its neuro-physiological structure. In ANN, computational units called neurons replace the nerve cells and the strengths of the interconnections are represented by weights, in which the learned information is stored. 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 multi-layered feed-forward network (MFFN) with multi-layered perceptron (MLP), mostly comprising three sequentially arranged layers of processing units (Jones and Hoskins, 1987; Girosi and Poggio, 1990). The MFFN provides a mapping between an input (x) and an output (y) through a nonlinear function f, i.e., $y = f(x)$.

2.1. Training Algorithm

The problem of neural network training is to obtain a set of interconnection weights such that the prediction error defined by the difference between the networks predicted outputs and the desired outputs is minimized. A steepest descent algorithm known as generalized delta rule (Baughman and Liu, 1995) is commonly used to modify the interconnection weights so as to minimize the prediction error. The iterative training makes the network to recognize patterns in the data and creates an internal model, which 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}:

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

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 non-decreasing function. The output layer of a single hidden layer network performs the same calculations as above, 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 of measuring 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. There after, weights are adjusted so as 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 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 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 to the output layer neuron k is computed as

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

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

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

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

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

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

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

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.

2.2. Information Processing

Network training is an iterative procedure that begins with initializing the weight matrix randomly. Network learning process involves a forward and a reverse pass. In the forward pass, an input pattern from the example data set is applied to the input nodes, the weighted sum of the inputs to the active node is calculated which is then transformed into output using a nonlinear activation function such as 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 Eq. (5) is computed and used for updating the network weights in accordance with the gradient strategy. 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.

3. Artificial Neural System for Mosquito Species Analysis

Artificial neural networks with back propagation 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, back propagation neural networks have been used to predict protein secondary and tertiary structures (Holley and Karplus, 1989; Kneller et al., 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 develop novel neural network methodology for accurate classification of mosquito species based on the genetic information of the ITS2 sequences of *Anopheles* genera.

3.1. Data Collection

The ribosomal DNA sequence (ITS2) data of 18 species of *Anopheles* genera is collected in fasta format from National Center for Biotechnological Information (NCBI) nucleotide database (www.ncbi.nlm.nih.gov/) and considered for computational experiment. These species include *Anopheles albitarsis*, *Anopheles maculipennis*, *Anopheles annulipes*, *Anopheles culicifacies*, *Anopheles farauti*, *Anopheles fluviatilis*, *Anopheles minimus*, *Anopheles epiroticus*, *Anopheles sundaicus*, *Anopheles benarrochi*, *Anopheles daciae*, *Anopheles messeae*, *Anopheles sacharovi*, *Anopheles atroparvus*, *Anopheles arabiensis*, *Anopheles sinensis*, *Anopheles anthropophagus*, and *Anopheles lesteri*. The length of the sequences varies in the range of 200–500.

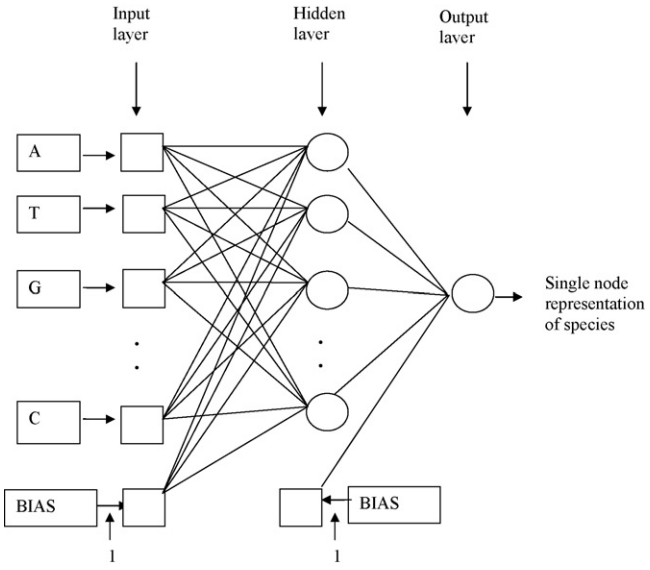

Fig. 1. MISONN structure.

3.2. 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 between the different classes of inputs. ITS2 ribosomal DNA sequences of mosquito species are made up of four bases, A, C, T, and G. These four bases should be represented in the form of a numerical vector in order to train and use a neural network for sequence classification. The numerical values to encode the input data sequences and to represent the output species are chosen so as to satisfy the training, prediction and classification requirements of the neural network configurations described in the following sections.

3.3. Multi-input Single-output Node Neural Network Model

The structure of MISONN is shown in Fig. 1. The bases A, C, T, and G of ribosomal DNA sequences form the network inputs with their corresponding species names as network outputs. The bases A, C, T, and G of the network inputs are coded with binary values and the network output species names are coded with real numbers. The bases in the network input sequences are assigned with A = {1 0 0 0}, T = {0 1 0 0}, G = {0 0 1 0} and C = {0 0 0 1}. This input assignment requires the number of active nodes in the input layer of the network as four times of the size of the sequence. The output coding employed to train the network is given in column 3 of Table 1. The coding in Table 1 is selected based on several trails of different input and output coding combinations and is found to provide effective network performance. A total of 18 species of *Anopheles* genera with each species having 10 data sequences of variable lengths are considered for neural network computations. Among these data sequences, six sequences of each species are used for training and the remaining four sequences are used for testing the trained network performance. For example, the ITS2 part of representative sample sequence of *A. benarrochi* used for training the network has the NCBI accession number AY684984. Similarly, the *A. arabiensis* used for testing the network has the NCBI accession number DQ287773.1. The input sequences used in training the

Table 1
Output coding of MISONN and MIMONN

S.No.	Species name	Output coding of "MISONN"	Output coding of "MIMONN"				
1	*Anopheles albitarsis*	10	0.3	0.0	0.0	0.0	0.0
2	*Anopheles anthrophagus*	20	0.0	0.3	0.0	0.0	0.0
3	*Anopheles arabiensis*	30	0.0	0.0	0.3	0.0	0.0
4	*Anopheles benarrochi*	40	0.0	0.0	0.0	0.3	0.0
5	*Anopheles daciae*	50	0.0	0.0	0.0	0.0	0.3
6	*Anopheles epiroticus*	60	0.45	0.0	0.0	0.0	0.0
7	*Anopheles farauti*	70	0.0	0.45	0.0	0.0	0.0
8	*Anopheles messeae*	80	0.0	0.0	0.45	0.0	0.0
9	*Anopheles minimus*	90	0.0	0.0	0.0	0.45	0.0
10	*Anopheles sacharovi*	100	0.0	0.0	0.0	0.0	0.45
11	*Anopheles sinensis*	110	0.6	0.0	0.0	0.0	0.0
12	*Anopheles sundaicus*	120	0.0	0.6	0.0	0.0	0.0
13	*Anopheles annulipes*	130	0.0	0.0	0.6	0.0	0.0
14	*Anopheles atroparvus*	140	0.0	0.0	0.0	0.6	0.0
15	*Anopheles culicifacies*	150	0.0	0.0	0.0	0.0	0.6
16	*Anopheles fluviatilis*	160	0.8	0.0	0.0	0.0	0.0
17	*Anopheles lesteri*	170	0.0	0.8	0.0	0.0	0.0
18	*Anopheles maculipennis*	180	0.0	0.0	0.8	0.0	0.0

network have different sizes. The network input nodes are to be chosen such that they suit to the sequence of any species irrespective of its size. In this work, a simple format is employed to manage the network with sequences of different sizes. According to this format, the number of input nodes to the input sequence of the network is specified with the binary coding equivalent to that of a larger sequence. This configuration facilitates the network to use the sequences of varying sizes by appropriately filling the additional nodes that exceed the sizes of shorter sequences. The network training is carried out by using back propagation delta rule algorithm. The network interconnection weights are initialized by assigning random numbers in the range of 0.2 to −0.2. 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 iterative convergence of the minimum objective makes the network to learn and modify the values of interconnection weights between the nodes of the layers. This minimization criterion also enables to select the number of hidden nodes and the number of iterations required for training. 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. During the association or classification phase, the trained neural network itself operates in a feed-forward manner. The number of hidden units, the number iterations, the learning rate α and the momentum factor η are selected in order to achieve the desired convergence in the objective function.

3.4. Multi-input Multi-output Node Neural Network Model

In MIMONN, the network input data of ribosomal DNA sequences are coded with binary values as in MISONN. The network outputs are specified with a binary encoding scheme given in column 4 onwards in Table 1. Back propagation delta rule algorithm is used to carry out the network training. All the data sets corresponding to input sequences with their binary coding and the output species with their corresponding coding are sequentially used to train the network model. In order to reduce the computational effort, the MIMONN is used with five output nodes corresponding to five species and the network is trained by sequentially feeding the outputs to these five nodes with their corresponding inputs. Output mapping 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 iterations, the learning rate α and the momentum factor η are selected in order to achieve the desired convergence in the objective function.

4. Results and Discussion

The number of hidden units, the number of iterations, the learning rate α and the momentum factor η are selected in order to achieve the desired convergence in the objective functions of both MISONN and MIMONN. Fig. 2 shows the iterative convergence of the error for MISONN with respect to the number of iterations for different hidden units. The parameters for both the networks are selected so that they provide better learning and generalization ability. The MISONN is configured by choosing the network parameters as $\alpha = 0.92$ and $\eta = 0.00179$ with 4 hidden nodes. The network has converged in 15000 iterations with a minimum training error of 0.84. The MIMONN is configured by setting the values of $\alpha = 0.80$ and $\eta = 0.00379$ with 20 hidden nodes. The network is converged in 20000 iterations with a minimum training error of 1.82. The trained and learned networks are then subjected to assess their recall and generalization abilities. The recall ability of the trained networks is evaluated by using the same input sequences as used for training. Both MISONN and MIMONN have exhibited 100% recall ability. The generalization ability of both the NN modeling configurations is evaluated by using the performance measures, namely, mean-squared error prediction (MSEP) and error prediction vari-

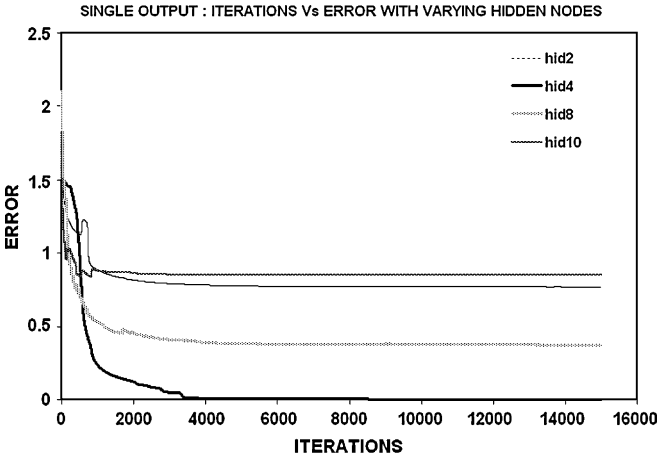

Fig. 2. Iterative error convergence in MISONN.

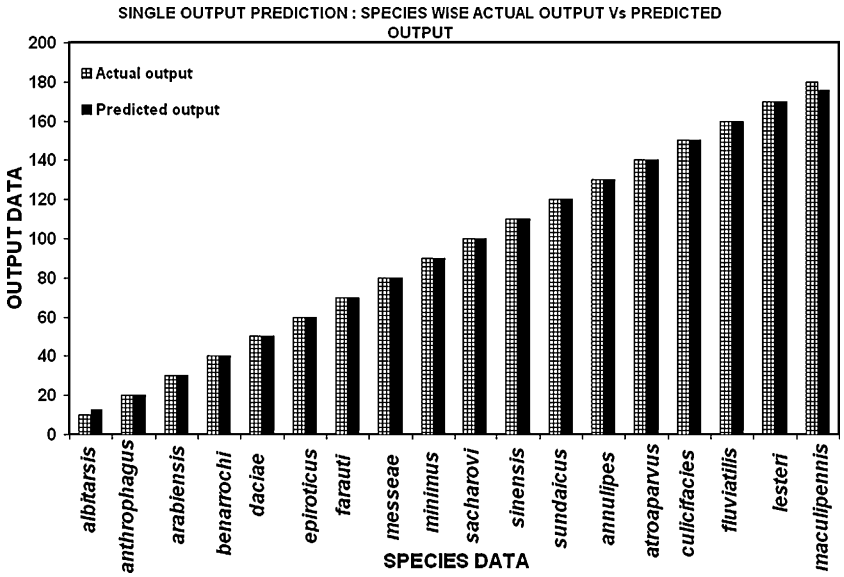

Fig. 3. Comparison of actual and predicted outputs of MISONN.

ance (EPV). The MSEP is defined by

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

where y_i and $\hat{y}_i$ are the target and predicted output values for the input sequences that are involved in training and n is the number data sequences used for prediction. The EPV in percent is defined as

$$\text{EPV} = \left\{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right\} \times 100 \tag{12}$$

where $\bar{y}$ is the mean value of target measurements used for prediction. The EPV measure on dividing by 100 gives the value of correlation coefficient (R^2).

The classification efficiency of MISONN and MIMONN is tested by using the species sequences that are not involved in training. When an untrained DNA sequence of a different species is used as input, the MISONN has exactly predicted the species corresponding to that sequence. Fig. 3 shows the network prediction results against their target values for each of the untrained sample sequences. The close agreement between the prediction results and the target values in Fig. 3 indicates the better classification efficiency of MISONN. The prediction results of MIMONN when tested with similar untrained species sequences are shown in Fig. 4. These results explain that the species predictions of some of the test sample sequences are not so close to their corresponding targets values. The results in Fig. 4 indicate the inferior classification efficiency of MIMONN over MISONN. The classification efficiency of MISONN and MIMONN are also evaluated in terms of the performance measures defined in Eqs. (11) and (12). Comparison of network models in terms of performance measures requires normalization of model outputs since they are in different magnitudes in both the models. Normalization of model outputs is performed by using a scaling measure of the form

$$y^s = \frac{y_j - \bar{y}}{\sigma_y} \tag{13}$$

Here y_j represents the actual or predicted species output, $\bar{y}$ is the mean output, σ_y is the standard deviation and y^s is the scaled output. The quantitative performance results have shown almost 100%

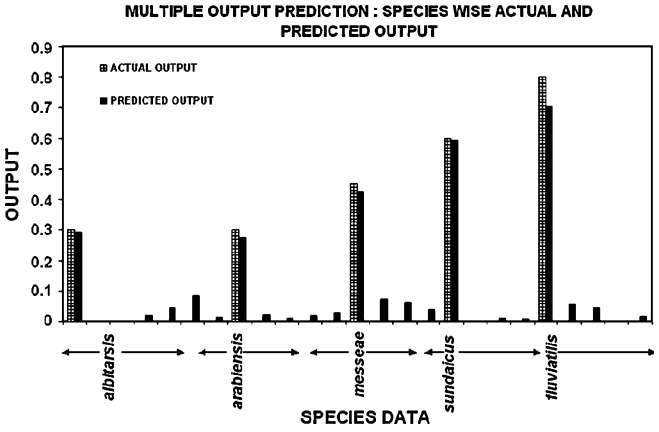

Fig. 4. Comparison of actual and predicted outputs of MIMONN.

recall ability of both the network models, but the generalization ability of the network models is more important for better classification of the species. The MSEP and EPV values for MISONN are evaluated as 0.00035477 and 99.962436, respectively. In the case of MIMONN, these values are found to be 0.2610 and 79.9680, respectively. The results thus demonstrate the better classification efficiency of MISONN over MIMONN.

5. Conclusions

Rapid and accurate identification and classification of mosquito species are of paramount importance for taking control measures against deadly diseases like malaria, filariasis, encephalitis, dengue fever and so on. Genetic identification is the confirmation for any kind of biological classification. The data sequences of ITS2 region considered for this work are widely used to extract the phylogenetic relation due to their well-conserved nature in a particular species. In this work, an artificial neural network method is presented for classification and identification of *Anopheles* mosquito species based on the genetic pattern information content of ITS2 ribosomal DNA sequences. The method is implemented in two different network model forms, namely, a multi-input single-output neural network a multi-input multi-output neural network. The perfor-

mance of both the network models for identification of *Anopheles* species is evaluated with respect to their prediction ability and generalization efficiency. The method of MISONN is found to provide accurate classification of mosquito species based on ITS2 ribosomal DNA sequences. The method presented in this study can be used for any kind of biological classification in genetic level.

References

Banerjee, A.K., Arora, N., Murty, U.S.N., 2007. Stability of ITS2 secondary structure in *Anopheles*: what lies beneath? Int. J. Integr. Biol. 1 (3), 232–238.

Baughman, D.R., Liu, Y.A., 1995. Neural Networks in Bioprocessing and Chemical Engineering. Academic Press, Dan Diego.

Blinder, P., Baruchi, I., Volman, V., Levine, H., Baranes, D., Jacob, E.B., 2005. Functional topology classification of biological computing networks. Nat. Comput. 4, 339–361.

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

Chen, B., Butlin, R.K., Pedrol, P.M., Wang, X.Z., Harbach, R.E., 2006. Molecular variation, systematics and distribution of the *Anopheles fluviatilis* complex in southern Asia. Med. Vet. Entomol. 20, 33–43.

Collins, F.H., Paskewitz, S.M., Finnerty, V., 1989. Ribosomal RNA genes of the *Anopheles gambiae* species complex. Adv. Dis.Vector Res. 6, 1–28.

Dayhoff, J.E., 1990. Neural Network Architectures. Van Nostrand Reinhold, New York.

Demeler, B., Zhou, G., 1991. Neural network optimization for *E. coli* promoter prediction. Nucleic Acids Res. 19, 1593–1599.

Dobzhansky, Th., 1937. Genetic nature of species differences. Am. Nat. 71 (735), 404–420.

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

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

Girosi, F., Poggio, T., 1990. Networks and the best approximation property. Biol. Cybern. 63, 169–179.

Holley, L.H., Karplus, M., 1989. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86, 152–156.

Jones, W.P., Hoskins, J., 1987. Back propagation. Byte, 155–162.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S.E., Sachs, J., 2004. A global index representing the stability of malaria transmission. Am. J. Trop. Med. Hyg. 70 (5), 486–498.

Kneller, D.G., Cohen, F.E., Langridge, R., 1990. Improvement in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214, 171–182.

Leonard, E.M.M., Jan, E.C., 1997. Systematics of mosquito disease vectors (Diptera, Culicidae): impact of molecular biology and cladistic analysis. Ann. Rev. Entomol. 42, 351–369.

Marinucci, M., Romi, R., Mancini, P., DiLuca, M., Severini, C., 1999. Phylogenetic relationships of seven Palearctic members of the maculipennis complex inferred from ITS2 sequence analysis. Insect Mol. Biol. 8 (4), 469–480.

Marrelli, M.T., Maria, A.M.S., Osvaldo, M., 2006. The second internal transcribed spacer of nuclear ribosomal DNA as a tool for Latin American anopheline taxonomy—a critical review. Mem. Inst. Oswaldo Cruz. 101 (8), 817–832.

Marrelli, M.T., Lucile, M.F.W., Malafronte, R.S., Tadei, W.P., Ricardo, L., Co-de-oliveira, Carmen, F.M., Osvaldo, M., 2005. Amazonian malaria vector anopheline relationships interpreted from ITS2 rDNA sequences. Med. Vet. Entomol. 19 (2), 208–218.

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

Sabbatini, R.M.E., 1993. Neural networks for classification and pattern recognition of biological signals. In: Proceedings of the 15th Annual International Conference of the IEEE, pp. 265–266.

Simpson, R.G., Williams, R., Ellis, R.E., Culverhouse, P.F., 1992. Biological pattern recognition by neural networks. Mar. Ecol. Prog. Ser. 79, 303–308.

Walton, C., Sharpe, R.G., Pritchard, S.J., Thelwell, N.J., Butlin, R.K., 1999a. Molecular identification of mosquito species. Biol. J. Linn. Soc. 68, 241–256.

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

Wilkerson, R.C., Reinert, J.F., Li, C., 2004. Ribosomal DNAITS2 sequences differentiate six species in the *Anopheles crucians* complex (Diptera: Culicidae). J. Med. Entomol. 41, 392–401.

Wu, C.G., Whitson, J., Mclarty, A.E., Chang, T.C., 1992. Protein classification artificial neural system. Protein Sci. 1 (5), 667–677.