VERSITA

Biologia, Bratislava, **62**/4: 462—469, 2007
Section Zoology
DOI: 10.2478/s11756-007-0089-1

Artificial neural networks for fly identification: A case study from the genera *Tachina* and *Ectophasia* (Diptera, Tachinidae)

Jaromír Vaňhara[1]*, Natália Murárikoá[1], Igor Malenovský[2] & Josef Havel[3]*

[1]*Masaryk University, Faculty of Science, Institute of Botany and Zoology, Kotlářská 2, CZ-61137 Brno, Czech Republic; e-mail: vanhara@sci.muni.cz*
[2]*Moravian Museum, Department of Entomology, Hviezdoslavova 29a, CZ-62700 Brno, Czech Republic*
[3]*Masaryk University, Faculty of Science, Department of Analytical Chemistry, Kotlářská 2, CZ-61137 Brno, Czech Republic; e-mail: havel@chemi.muni.cz*

Abstract: The classification methodology based on morphometric data and supervised artificial neural networks (ANN) was tested on five fly species of the parasitoid genera *Tachina* and *Ectophasia* (Diptera, Tachinidae). Objects were initially photographed, then digitalized; consequently the picture was scaled and measured by means of an image analyser. The 16 variables used for classification included length of different wing veins or their parts and width of antennal segments. The sex was found to have some influence on the data and was included in the study as another input variable. Better and reliable classification was obtained when data from both the right and left wings were entered, the data from one wing were however found to be sufficient. The prediction success (correct identification of unknown test samples) varied from 88 to 100% throughout the study depending especially on the number of specimens in the training set. Classification of the studied Diptera species using ANN is possible assuming a sufficiently high number (tens) of specimens of each species is available for the ANN training. The methodology proposed is quite general and can be applied for all biological objects where it is possible to define adequate diagnostic characters and create the appropriate database.

Key words: artificial neural networks; species identification; Diptera; Tachinidae; *Tachina*; *Ectophasia*; parasitoids

Introduction

In the last century, artificial neural network (ANN) computation was developed having been inspired by neurobiology and the way the human brain works. Nowadays, numerous and wide applications are found in all branches of science. There are extensive applications of ANN in chemistry, biochemistry as well as in biology. ANN is used, e.g., in microbiology, ecology, hydrobiology, entomology, and even in forensic investigations.

The use of ANN in taxonomy is rather rare, even though in this discipline they clearly provide a powerful pattern recognition and data analysis tool as pointed out already by Weeks & Gaston (1997). Their ability to learn patterns in multivariate data makes them ideally suited to identification problems.

Identification of species has always been an object of general interest and often an important task in systematic biological disciplines, such as botany, zoology, as well as entomology. Up to now, the traditional methods for the identification of species are mostly based on great encyclopaedic knowledge, memory, and many long-years experience of an expert in the field. ANN, however, have a great potential to partly automate the identification process, especially if coupled with image analysis (Weeks & Gaston 1997; Gaston & O'Neill 2004). This was also documented in the case study by Do et al. (1999) who successfully applied ANN to identify six spider species from the family Lycosidae using transformed digital images of female genitalia. The ANN have also performed the treatment of bioacoustic data, e.g., Chesmore (2004) used them for automated identification of four species of British Orthoptera from sound recordings. Hernández-Borges et al. (2004) identified three species of limpets (Mollusca, Patellogastropoda, *Patella*) in the coastal areas of the Canary Islands using ANN and chemotaxonomic data (the content of different aliphatic hydrocarbons).

Quite often, in some taxonomic groups the characters used for the discrimination of species are not unambiguously distinct. They may overlap, be variable or even missing and the species might then be difficult to separate. Studying the flowering plant genus *Lithops* (Aizoaceae), Clark (2003) suggested that, in such cases, ANN can be used as advisory tools for taxonomists. This is particularly useful since experts in different groups often work alone and an unbiased second opinion on the identification of a "difficult" specimen is valuable. Frequently, the difficulties in the identifi-

* Corresponding authors: J. Vaňhara (Diptera), J. Havel (ANN)

cation pertain to some developmental stage or sex. For example, the Neotropical Phlebotomine sand fly species of the *Lutzomyia intermedia* (Lutz & Neiva, 1912) complex (Diptera, Psychodidae) can be easily distinguished by females but their males are very similar. Marcondes & Borges (2000) succeeded in reliably distinguishing the males using ANN and a set of morphometric characters and thus brought the first application of ANN in Diptera.

One of the most taxonomically problematic families of Diptera are the Tachinidae (McAlpine 1989), also being one of the largest fly families, with about 8,200 species worldwide. Among the species 1,550 occur in the Palaearctic Region and nearly 880 in Europe (Tschorsnig & Richter 1998; Tschorsnig et al. 2005). Tachinids are parasitoids of insects and some other arthropods, and may be regarded as economically beneficial as they also prey on many agricultural and forest pests. The Central European fauna was treated in the monograph by Tschorsnig & Herting (1994). Some persistent doubts and taxonomic problems concern, e.g., the genera *Tachina* and *Ectophasia*, due to a great morphological variability of individual species in Central Europe (Vaňhara et al. 2004). Often, this causes an uncertain and ambiguous identification of some specimens, which can be otherwise solved out only by much experience and a good reference collection.

The aim of our study is to select an alternative set of diagnostic characters and evaluate the possibility of supervised ANN to identify individual specimens in these two model genera of Tachinidae.

Theory of ANN

ANN represent sophisticated computational modelling tools which can be used to solve a wide variety of complex problems. The attractiveness of ANN in biology comes from their capability to learn and/or model very complex systems which allows for the possibility of them being used as a tool for classification. Therefore, their actual potential in this branch of science is high. A big contrast between conventional computer programs and ANN is reflected in the fact that the former can only accomplish those tasks for which they were specifically designed, while ANN is a kind of generalised learning machine which can, in principle, learn almost anything. ANN's theory has been widely discussed in previous literature. Good overview of ANN principles can be found in the monographs by Carling (1992), Fausett (1994), Bishop (1995), Patterson (1996); the theory of different networks has been reviewed by Zupan & Gasteiger (1991). Here we will only briefly describe the basic idea of this methodology.

ANN is a computational model formed from a certain number of single units, artificial neurones or nodes, connected with coefficients (weights), $w_{ij,}$ which constitute the neural structure. Many different neural network architectures can be used. One of the most common is the feed forward supervised neural network of multi-layer perceptrons (MLP). The MLP is conventionally constructed with three or more layers, i.e., input, output and hidden layers (Fig. 1).

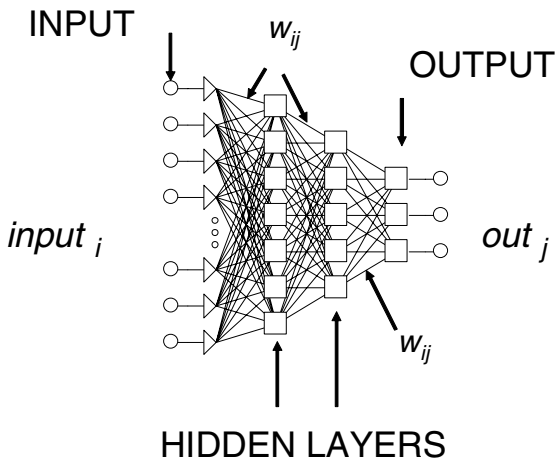

Fig. 1. Construction of ANN from four layers, i.e., input, output and two hidden layers.

Each layer has a different number of nodes. The input layer receives the information about the system (the nodes of this layer are simple distributive nodes which do not alter the input value at all). The hidden layer processes the information initiated at the input, while the output layer is the observable response or behaviour. The inputs, $input_i$, multiplied by connection weights $w_{ij,}$ are first summed and then passed through a transfer function to produce the output, out_i. The determination of the appropriate number of hidden layers and number of hidden nodes in each layer is one of the most critical tasks in ANN design. Unlike the input and output layers, one starts with no prior knowledge of the number and size of hidden layers.

The use of ANN consists of two steps: "Training" and "Prediction". The "Training" consists first of defining input and output data to the network. It is usually necessary to scale the data or normalize it to the networks paradigm. This data is referred to as the training set. In this training phase, where actual data must be used, the optimum structure, weight coefficients and biases of the network are searched for. The training is considered complete when the neural networks achieved the desired statistical accuracy as it produces the required outputs for a given sequence of inputs. A good criterion to find the correct network structure and therefore to stop the learning process is to minimize the root mean square error (RMS)

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - out_{ij})^2}{N \times M}}$$

where y_{ij} is the element of the matrix $(N \times M)$ for the training set or test set, and out_{ij} is the element of the output matrix $(N \times M)$ of the neural network, where N is the number of variables in the matrix and M is the number of samples. RMS gives a single number which summarizes the overall error.

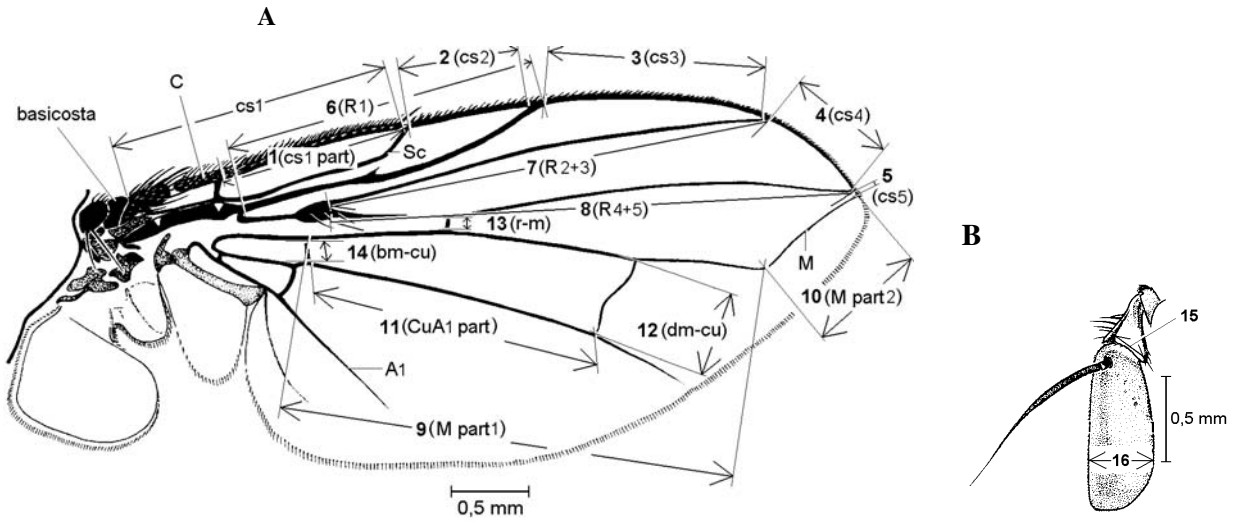

Fig. 2. Measured wing (A 1–14) and antennal (B 15–16) characters in Tachinidae.
1(cs_1 part) – part of costal section cs_1; 2(cs_2) – length of costal section cs_2; 3(cs_3) – costal section cs_3; 4(cs_4) – cs_4; 5(cs_5) – cs_5; 6(R_1) – length of R_1; 7(R_{2+3}) – length of R_{2+3}; 8(R_{4+5}) – R_{4+5}; 9(M part1) – medial vein M between bm-cu and its bend; 10(M part2) – length of post angular vein (the rest of M); 11(CuA_1 part) – CuA_1 between bm-cu and dm-cu; 12(dm-cu) – length of dm-cu; 13(r-m) – length of r-m; 14(bm-cu) – length of bm-cu. 15(2^{nd}) – width of antennal segment 2; 16(3^{rd}) – width of antennal segment 3.

After a supervised network performs well on training data, it is important to check what it can do with the data it has not seen before. This is called verification. This testing is critical to insure that the network has not simply memorized the training set but has learned the general patterns involved within an application. At this stage other input data are submitted to the network in order to evaluate if it can predict the outputs. In this case the outputs are already known, but they are not shown to the network. The predicted value is compared to the experimental one to see how well the network is performing. If the system does not give reasonable outputs for this test set, the training period is not over or the network is able to model the data but cannot predict them. When the verification of the known data is passed well, in the final stage, unknown data is evaluated and the outputs are predicted (classification of unknown data is done).

Material and methods

A group of several model species selected from the parasitoid family Tachinidae (Diptera) was used for ANN studies, namely three species of the genus *Tachina* Meigen, 1803: *T. fera* (L., 1761), *T. magnicornis* (Zetterstedt, 1844), *T. nupta* (Rondani, 1859), and two species of the genus *Ectophasia* Townsend, 1912: *E. crassipennis* (F., 1794) and *E. oblonga* (Robineau-Desvoidy, 1830). The material included altogether 113 specimens (58 males and 55 females) which were a priori identified or revised by J. Vaňhara using the key of Tschorsnig & Herting (1994).

Dry-mounted (pinned) adults were initially photographed (stereomicroscope Olympus SZX 12 with attached Camedia C-5050 digital camera), consequently the digitalized picture was scaled (in μm) by means of an image analyser using software M.I.S QuickPhoto Micro Olympus (Japan).

The wing shape and vein pattern is at once quite conservative within different insect groups and almost all taxa are diagnosably different from one another. This seems to be true for insects in general (Kukalová-Peck 1991) and Diptera in particular (e.g., Hennig 1954; Houle et al. 2003). Moreover, the transparent, two-dimensional structure and obvious venation make insect wings ideal subject for image analysis in taxonomy (Weeks & Gaston 1997; Weeks et al. 1997). We preferentially selected 14 morphometric characters on wings, which included length of different wing veins or their parts (Fig. 2A). In *Tachina* only one wing was measured, in *Ectophasia* both the right and left wing data were recorded, if possible, for each specimen. Furthermore (in *Tachina* only), we measured the width of antennal segments 2 and 3 (Fig. 2B). All these linear distances were measured by manually pointing landmarks with a mouse on computer screen in the QuickPhoto program which then automatically provided the value of the measurement. Sex of the studied specimen was recorded as another variable. Thus, there were altogether 17 variables in *Tachina* (14 for one of the wings, two for antenna, and one for sex) and 29 variables (14 for left, 14 for right wing, and one for sex) in *Ectophasia*.
Character abbreviations used:
RW, LW – right and left wings
1-(cs_1 part) – length of the part of section cs_1 of costa C;
2-(cs_2) – length of costal section cs_2;
3-(cs_3) – length of costal section cs_3;
4-(cs_4) – length of costal section cs_4;
5-(cs_5) – length of costal section cs_5;
6-(R_1) – length of radius R_1;
7-(R_{2+3}) – length of radius R_{2+3};
8-(R_{4+5}) – length of radius R_{4+5};
9-(M part1) – length of the basal part of medial vein, i.e. between cross-vein bm-cu and bend of M;
10-(M part2) – length of post angular vein, i.e. medial vein M from its bend to the end;
11-(CuA_1 part) – length of anterior branch CuA_1 of cubital vein, its part between bm-cu and dm-cu;
12-(dm-cu) – length of discal medial-cubital cross-vein dm-cu;
13-(r-m) – length of radial-medial cross-vein r-m;
14-(bm-cu) – length of basal medial-cubital cross-vein bm-cu;

15-(2^{nd}) – the width of antennal segment 2 in its widest part, in *Tachina* only;
16-(3^{rd}) – the width of antennal segment 3 in its widest part, in *Tachina* only;
17-(sex M, F) – male, female.

ANN computation and program

ANN computation was performed using TRAJAN Neural Network Simulator, Release 3.0 D. (TRAJAN Software Ltd 1996–1998, UK). Some calculations (summary statistics, *t*-test for dependent samples, cluster analysis) were also done using STATISTICA V.7 (StatSoft, Inc., USA). All computation was performed on a standard PC computer with operating system Microsoft Windows Professional XP.

Remark about the computational strategy and ANN procedures

Data for ANN computing were randomly divided into: 1. Learning set; 2. Verification set; and 3. Test set.

The learning set consists of a number of samples (specimens of flies a priori identified to species) characterized by variables (characters) obtained by image analysis. This set is used to search for the suitable architecture of ANN by which the classification is made possible. The process which is based on searching for corresponding weights w_{ij} in order to minimize the RMS value is called learning or training. The correctness of the model obtained using the most suitable data set, architecture and training procedure is then verified on another independent set of samples called the verification set. Finally, the obtained ANN model can be used to classify principally unknown specimens (Test set).

Each set represents a matrix with the number of columns corresponding to the number of variables. The variables, in the input and/or output, can have numerical values, but also nominal (categorial, e.g., male/female, the name of the species).

Missing values

It can happen that the matrix cannot be completed because some values of variables within data sets are not known or cannot be obtained (damage in one of the wings, etc.). Although such cases that contain missing data are problematic, they can still be used in data analysis. There are various methods to manage with missing data (e.g., by mean substitution, various types of interpolations and extrapolations).

Results and discussion

Case (1): Identification of the model species of the genus Tachina

In this Case (1) in order to examine the possibilities of ANN for the classification of 75 fly samples, a total of 17 characters were used for the species *T. fera* (49), *T. magnicornis* (20) and *T. nupta* (6).

Because the number of specimens for *T. nupta* was limited, we first examined the possibility of only classifying the first two rich species, i.e., *T. fera* and *T. magnicornis,* where we had a higher number of specimens.

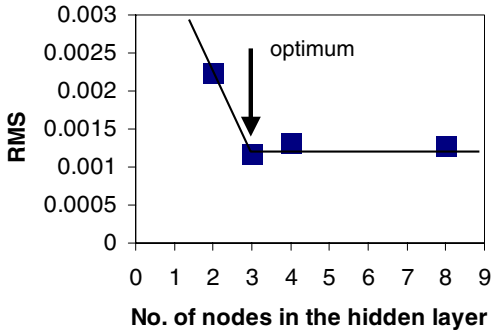

Fig. 3. Search of the optimal architecture for the classification of two species, i.e., *Tachina fera* and *T. magnicornis*.

Classification of T. fera and T. magnicornis

There were 49 specimens of *T. fera* and 20 of *T. magnicornis* available. In the first stage the optimal architecture (17, *n*, 2) for the modelling was investigated. The number of inputs (17) is the number of nodes in the input layer, *n* the number of nodes in the hidden layer, and 2 the number of nodes in the output layer (the name of the species). Plotting the RMS value as a function of the number of nodes in the hidden layer, as shown in Fig. 3, the optimal number of nodes in the hidden layer, *n*, was equal to 3. Thus the modelling with ANN was done using the ANN architecture (17, 3, 2) and it was found that all the samples from the training set were correctly classified in the learning process.

In order to prove the classification and prediction power of ANN, some of the samples were arbitrarily excluded from the training set (the leave-one-out methodology) and then used as the verification samples. For just one verified specimen 100% correct classification was always reached; for up to 5 specimens excluded from the training set, the verification (prediction) was from 98.6 to 100%, and for 10 samples 94.2–100%. The decrease in the correctness of prediction is understandable as the number of specimens used in the learning set was also decreasing. Anyway, it can be concluded that the variables selected in the case under study are representative and contain sufficient resolving power for correct ANN classification and prediction. Also the number of samples in the training set was sufficient. Generally, the greater the number of specimens, the better the prediction that can be expected.

The role of the sex variable in input data

The information on specimen's sex was also given as a variable in the input layer. However, the problem was if the sex would be considered so important as to be included in the analysis. When the variable sex was eliminated from the learning set, the classification of *T. fera* and *T. magnicornis* hardly changed. However, the information on sex remains implicitly in the data (as specimens of different sex were mixed together). When the classification (name) of the specimen was used as a variable in the input layer and the sex as an output variable, almost all the samples were correctly classified to the assigned sex in the training process.

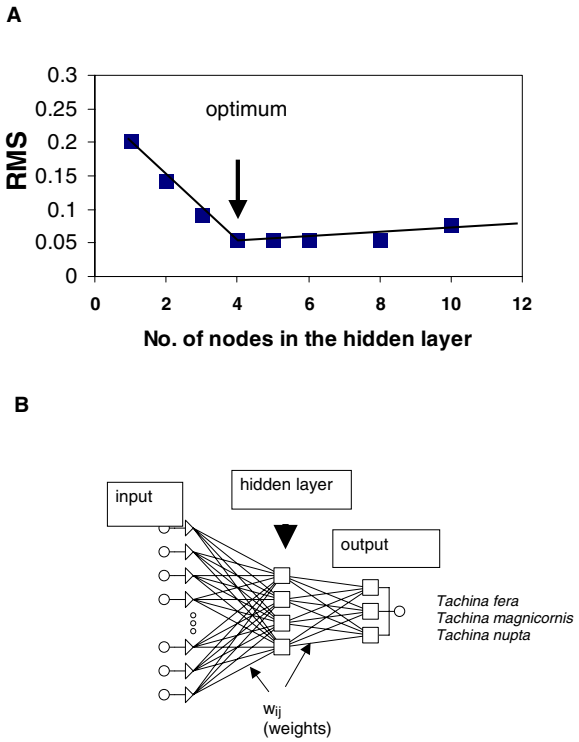

Fig. 4. Search of the optimal architecture (A) and an example of the optimal Artificial Neural Networks architecture (B) for classification of 3 spp., i.e., *Tachina fera, T. magnicornis* and *T. nupta.*

It is therefore evident that the variable, sex, has some value. Therefore, in the next step, only data corresponding to specimens with sex = males (M) was used in the training set (and at the same time the variable sex was NOT given in the input). Using such input data in the learning set the classification of females (F, output variable) was done using the same neural architecture as before. Complete 100% prediction was obtained only for *T. fera*, while most of *T. magnicornis* were either classified wrongly or classified as unknown.

In conclusion, there is almost no difference between male and female specimens for *T. fera*, but some sexual difference is observed between specimens of *T. magnicornis*. Therefore, in the next study the variable sex was kept in the input data and used as one of the input variables.

Classification of T. fera (49), T. magnicornis (20), and T. nupta (6), i.e., 75 specimens

Even if the number of samples for one of the species was quite low (6 samples), it was examined to see whether *T. nupta* can be classified and/or distinguished from *T. fera* and *T. magnicornis* at all. Following the same procedure as in the cases described above, the result of the search of optimal architecture is shown in Fig. 4A. Evidently, a more complex architecture is needed, where the number of nodes in the hidden layers is equal to 4. The corresponding architecture (17, 4, 3) is present in Fig. 4B.

A traditional cluster analysis, cf. Fig. 5 demonstrated that, in this case, the classification is very difficult and complicated. Using ANN for direct clustering (classification) is however more straightforward.

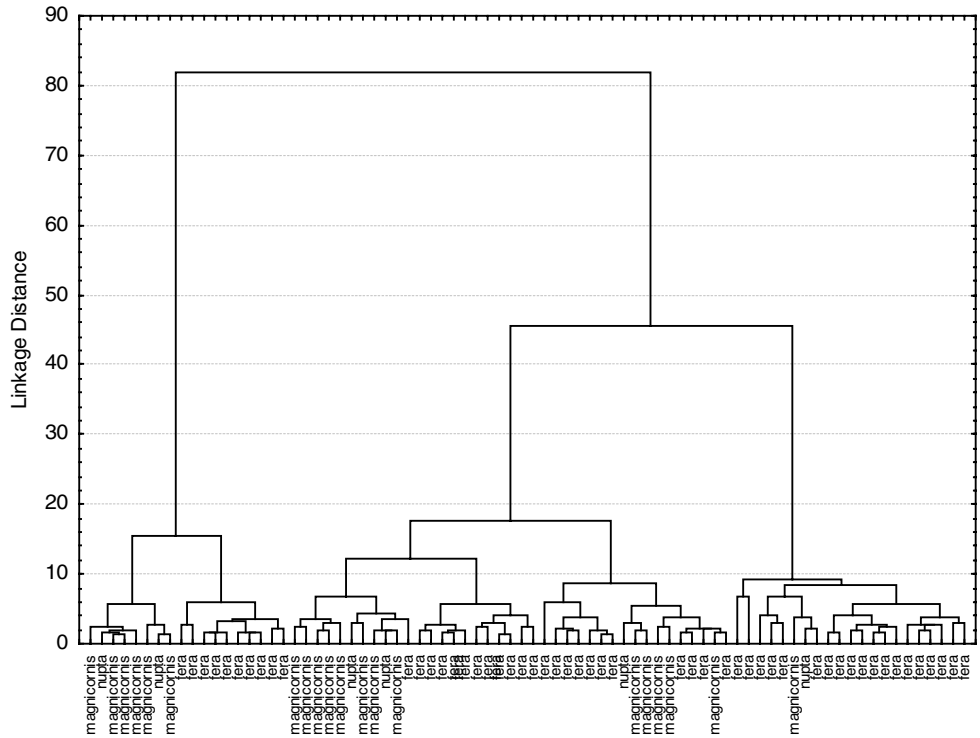

Fig. 5. Classification of three species of *Tachina* using cluster analysis (Ward's method, Euclidean distance metric).

Application of the whole data for the species classification

The complete data of all samples including all species, i.e., *T. fera* (49), *T. magnicornis* (20), and *T. nupta* (6) and the architecture (17, 4, 3) were examined with respect to the prediction of unknown samples. All samples from *T. fera* and *T. magnicornis* were classified correctly. In the rare *T. nupta,* single specimens were either wrongly classified or classified as unknown. Summarizing, 95–100% prediction was reached. It is suggested that better and more reliable prediction could be reached by increasing the number of samples in the learning set for *T. nupta*.

Conclusions from Case 1

We have found that the optimal ANN architecture is a three-layer ANN with the number of nodes in the hidden layer equal to a max. of four. This eliminated "over-training" and thus ensured reliable prediction.

Classification of *T. fera, T. magnicornis*, and *T. nupta* by ANN is possible. In order to reach a high degree of certainty and reliability of the prediction, the number of specimens of *T. nupta* should be enlarged to be comparable with the first two species. In the biometric database the sex should be given as one of the input characters as well.

Case (2): Recognition of Ectophasia crassipennis and E. oblonga

E. crassipennis and *E. oblonga* data from the LEFT (14 morphometric characters) and RIGHT (also 14) wings were collected from 25 specimens of *E. crassipennis* and 13 specimens of *E. oblonga*.

In the first stage all data (including left and right wings all together) was used to search for the ANN architecture and training, in spite of the fact that some of the data was missing (12 % of entries pertaining to damaged specimens). The standard procedure to replace missing data as offered in the Trajan program was applied.

The optimal architecture of ANN for training was found to be (38, 4, 2) and, using it, the training was successful. All samples were correctly classified. Verification of the model was done by selecting arbitrarily by a Monte Carlo method from 1 up to 5 specimens and the prediction was always correct. Even when 10 specimens were excluded from the training set (moving them to the Verification set), only 1–2 specimens were not correctly classified or they were considered as unknowns. It can be concluded that the variables are adequately selected and enable correct classification as well as excellent prediction.

Examination of the sex variable

When the sex variable was removed from the input there was no apparent effect. All the samples were again classified correctly. In spite of this, the significance of this variable was studied further.

Because the number of samples was rather low, an attempt was made to predict the sex variable from all the input data, including species classification, in the learning set. It was found that all samples were correctly classified according to sex. Even testing the prediction of sex for the 1–3 samples excluded from the process of training was successful (from 95 to 100%). It seems therefore that sex has some effect on the other input parameters and so whenever possible this variable should be given as an input.

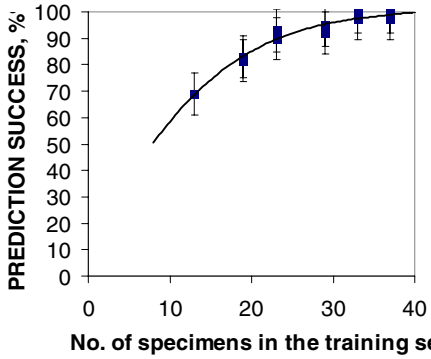

Fig. 6. Success of the prediction concerning classification of *Ectophasia crassipennis* and *E. oblonga* (input data from the Right wing only were used, all data from the Left wing were eliminated).

The effect of the right or left wing variables

The important question is whether the data from the left wing is the same as from the right one and also if the data from the left and right wings is interchangeable. In order to find the influence of the left/right on the classification process, the data from the right wing was used solely in the training phase and it was demonstrated that just data from the right wing only can be used for the classification because all of the specimens were classified correctly in the learning phase. Then, the verification was examined. The data in the training set was systematically reduced and more and more samples were excluded from the Training set by a Monte Carlo method and given to the Verification set.

Fig. 6 shows how the success of the prediction depends on the number of samples. It follows from this figure that for a low number of samples in the verification set (i.e., no. of samples in the training set was not too diminished); the prediction ability does not change much. Keeping the number of samples to at least 30 or more, the success of prediction is higher than 98% (calculated with respect to the total no. of samples).

Similar results were obtained for the ANN classification from the left wing measurements (results not given here).

Classification separating the data for the right and left wings

We have also examined the use of ANN when data from the right and left wings were analyzed separately. Each specimen of *Ectophasia* was now entered twice in the data set, once with the right wing values only and once with the left wing data. Data evaluation gave the fol-

Table 1. Comparison of right and left wing data in *Ectophasia*.

Character no.	Right wing		Left wing		t-test for dependent samples		
	mean	S.D.	mean	S.D.	n	t	P
1	1833.06	391.71	1877.34	388.62	28	**−3.04**	**0.01**
2	959.29	207.13	990.22	211.27	30	**−2.90**	**0.01**
3	1894.08	330.09	1900.78	336.53	30	−0.45	0.65
4	701.23	116.77	690.37	120.37	30	1.46	0.16
5	126.24	30.94	124.28	31.87	29	0.58	0.57
6	2936.27	602.28	3029.63	582.08	30	−1.91	0.07
7	4179.47	759.06	4194.68	729.98	30	−0.38	0.70
8	4742.33	809.34	4698.63	790.36	30	1.35	0.19
9	3668.37	639.15	3668.37	628.62	30	0.00	1.00
10	1517.96	270.07	1442.35	215.79	26	**3.67**	**0.001**
11	2498.66	520.46	2390.89	467.40	30	**2.60**	**0.01**
12	1118.01	257.07	1181.83	257.64	27	**−3.27**	**0.003**
13	277.30	81.44	282.94	86.19	28	−0.92	0.36
14	228.04	59.05	219.51	45.47	29	1.39	0.18

Explanations: Mean and standard deviation are given in μm. Boldface values are significant at $P < 0.05$.

lowing results: 43 out of 50 wings of *E. crassipennis* and 24 out of 26 of *E. oblonga* were correctly classified. A single wing of *E. crassipennis* was classified wrongly, six wings of *E. crassipennis* and 2 of *E. oblonga* were classified as unknown (unclassified). These results are slightly worse than when each specimen was characterized by data from both the right and left wings.

Conclusion of Case 2

Because the whole collection of the samples (38 in total) is rather small, it is difficult to decide if there are statistical differences between the Right and Left wings. Using the t-test for dependent samples, differences significant at $P < 0.05$ between the right and left wings were found in five out of the 14 characters (Tab. 1). The use of both the right and left wing data for each specimen probably brings some dispersion which might reflect the real variation in fly morphology and the increase of the number of data in input improves the prediction. However, a more decisive answer can always be obtained by increasing the number of samples.

Conclusions and perspectives

The use of Artificial Neural Networks for species identification of Diptera was tested on model species from the family Tachinidae (3 spp. of *Tachina,* 2 spp. of *Ectophasia*) which are difficult to identify by conventional taxonomic procedures. Biometric input data measured both in males and females took in 16 characters of one wing and antenna (*Tachina*), 14 characters of both the right and left wings (*Ectophasia*), and the sex of the studied specimen. It was found that the variables selected in this study are representative and contain sufficient resolving power for correct ANN classification and prediction, assuming the number of samples in the training set is sufficiently high. Generally, the higher the number of specimens, the better the prediction observed. Data in all cases studied was not only successfully modelled by ANN but also it was found possible to predict and identify the new unknown samples with excellent success.

Generally, the methodology based on ANN can be applied for any other biological objects where it is possible to define adequate diagnostic characters.

In conclusion, with a sufficiently extensive and reliable database, the Artificial Neural Networks represent a new powerful tool for the fast and objective identification of flies even in taxonomically difficult groups and open new possibilities for taxonomy. In some cases, they could offer an alternative to molecular diagnostic methods which are becoming widely used in current entomology (e.g., Tóthová et al. 2006). Comparing to those, ANN coupled with image analysis is a relatively cheap and non-destructive analytic method which can thus be used also, e.g., for studying type material or collection material permanently mounted on slides. Moreover, the identification process can be considerably speeded up with automating the image analysis system (Houle et al. 2003; Tofilski 2004).

Acknowledgements

For financial support, the Ministry of Education and the Masaryk University (grant No. MSM 0021622416) and Grant Agency of the Czech Republic (grant No. 524/05/H536) are acknowledged. The paper was also partly supported by the grant (No. MK 00009486201) from the Ministry of Culture of the Czech Republic to the Moravian Museum. English was kindly revised by Mr. Phil G. Watson.

References

Bishop C. 1995. Neural Networks for Pattern Recognition. Oxford University Press, 504 pp.

Carling A. 1992. Introducing Neural Networks. Sigma Press, Wilmslow, UK, 338 pp.

Clark J.Y. 2003. Artificial neural networks for species identification by taxonomists. BioSystems **72:** 131–147.

Chesmore D. 2004. Automated bioacoustic identification of species. An. Acad. Bras. Cienc. **76:** 435–440.

Do M.T., Harp J.M. & Norris K.C. 1999. A test of a pattern recognition system for identification of spiders. Bull. Entomol. Res. **89:** 217–224.

Fausett L. 1994. Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall, New York, 461 pp.

Gaston K.J. & O'Neill M.A. 2004. Automated species identification: why not? Philos. Trans. R. Soc. Lond. B Biol. Sci. **359:** 655–667.

Hennig W. 1954. Flügelgeäder und System der Dipteren unter Berücksichtigung der aus dem Mesosoikum beschriebenen Fossilien. Beitr. Entomol. **4:** 245–388.

Hernández-Borges J., Corbella-Tena R., Rodriguez-Delgado M.A., Garcia-Montelongo F.J. & Havel J. 2004. Content of aliphatic hydrocarbons in limpets as a new way for classification of species using Artificial Neural Networks. Chemosphere **54:** 1059–1069.

Houle D., Mezey J., Galpern P. & Carter A. 2003. Automated measurement of *Drosophila* wings. BMC Evolutionary Biology **3:** 25. doi:10.1186/1471-2148-3-25. http://www.biomedcentral.com/1471-2148/3/25 (accessed 12.10.2006).

Kukalová-Peck J. 1991. Fossil history and the evolution of hexapod structures, pp. 141–179. In: Naumann I.D. (ed.), The Insects of Australia, Vol. 1, Melbourne Univ. Press.

McAlpine J.F. 1989. Phylogeny and classification of the Muscomorpha, pp. 1397–1505. In: McAlpine J.F. (ed.), Manual of Nearctic Diptera, Vol. 3, Monograph No. 32, Research Branch, Agriculture Canada, Ottawa.

Marcondes C.B. & Borges P.S. 2000. Distinction of males of the *Lutzomyia intermedia* (Lutz & Neiva, 1912) species complex by ratios between dimensions and by an Artificial Neural Network (Diptera: Psychodidae, Phlebotominae). Mem. Inst. Oswaldo Cruz **95:** 685–688.

Patterson D. 1996. Artificial Neural Networks: Theory and Applications. Prentice Hall, Singapore.

Tofilski A. 2004. DrawWing, a program for numerical description of insect wings. J. Insect Sci. **4:** 17. http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=528877& blobtype=pdf (accessed 12.10.2006).

Tóthová A., Bryja J., Bejdák P. & Vaňhara J. 2006. Molecular markers used in phylogenetic studies of Diptera with a methodological overview. Dipterologica Bohemoslovaca, Vol. 13, Acta Univ. Carol. Biol. **50:** 125–133.

Tschorsnig H.-P. & Herting B. 1994. Die Raupenfliegen (Diptera: Tachinidae) Mitteleuropas: Bestimmungstabellen und Angaben zur Verbreitung und Ökologie der einzelnen Arten. Stuttgarter Beiträge zur Naturkunde, Serie A, No. **506**, pp. 1–170. Online authorized version of English translation by Rayner R. & Raper C. "Tschorsnig H.-P. & Herting B. 2001. The Tachinids (Diptera: Tachinidae) of Central Europe: Identification keys for the species and data on distribution and ecology". http://tachinidae.org.uk/site/downloads.php (accessed 12.10.2006).

Tschorsnig H.-P. & Richter V. 1998. Tachinidae, pp. 691–827. In: Papp L. & Darvas B. (eds), Contributions to a Manual of Palaearctic Diptera (with special reference to flies of economic importance), Vol. 3., Higher Brachycera, Science Herald, Budapest.

Tschorsnig H.-P., Richter V.A., Cerretti P., Zeegers T., Bergström C., Vaňhara J., Van de Weyer G., Bystrowski C., Raper C., Ziegler J. & Hubenov, Z. 2005. Tachinidae. In: Fauna Europaea Service, Version 1.2. http://www.faunaeur.org (accessed 12.10.2006).

Vaňhara J., Tschorsnig H.-P. & Barták M. 2004. New records of Tachinidae (Diptera) from the Czech Republic and Slovakia, with revised check-list. Stud. Dipterol. **10 (2003):** 679–701.

Weeks P.J.D. & Gaston K.J. 1997. Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. Biodivers. Conserv. **6:** 263–274.

Weeks P.J.D., Gauld I.D., Gaston K.J. & O'Neill M.A. 1997. Automating the identification of insects: a new solution to an old problem. Bull. Entomol. Res. **87:** 203–211.

Zupan J. & Gasteiger J. 1999. Neural Networks in Chemistry and Drug Design. Wiley-VCH, Weinheim, 380 pp.

Received October 27, 2006
Accepted November 27, 2006