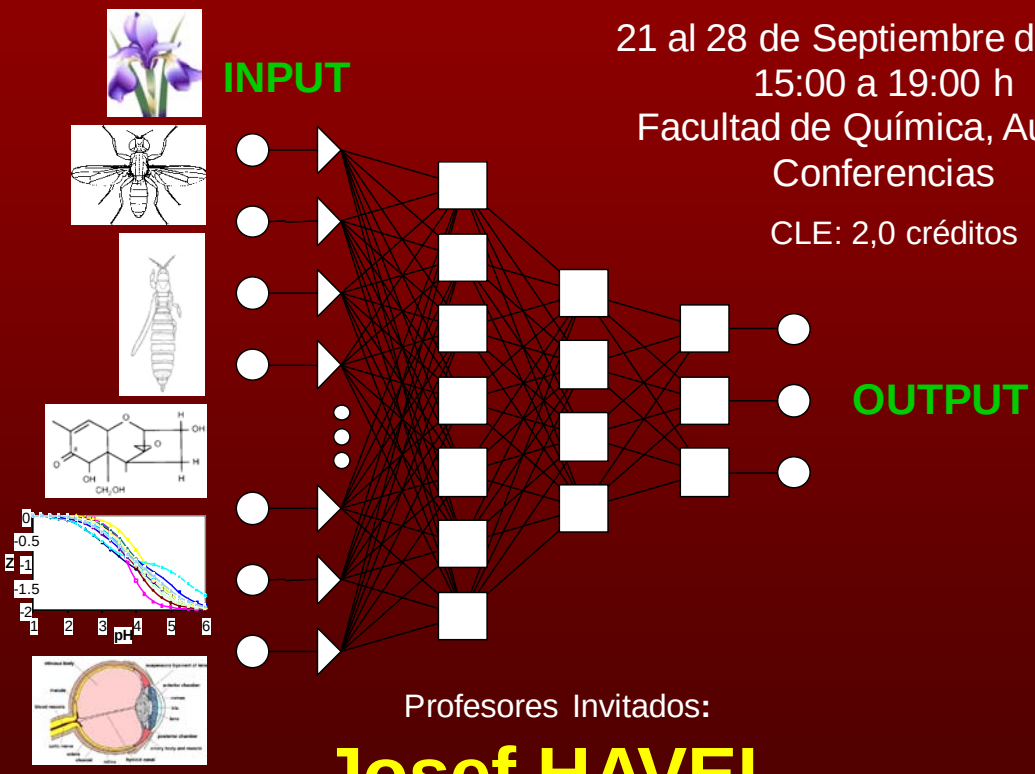
Aplicaciones de Redes Neuronales (ANN) en Ciencia

Facultad de Biología & Facultad de Química,

UNIVERSIDAD de LA LAGUNA, Tenerife, España



21 al 28 de Septiembre de 2009
15:00 a 19:00 h
Facultad de Química, Aula de Conferencias

CLE: 2,0 créditos

Profesores Invitados:

Josef HAVEL
Jaromír VAŇHARA

Departamento de Química; Departamento de Botánica y Zoología,
Facultad de Ciencias, Universidad de Masaryk, Brno, República Checa

Directores:

Dr. M. BÁEZ
Departamento de Zoología
Facultad de Biología
marbaez@ull.es

Dr. E.M. PEÑA-MÉNDEZ
Departamento de Química Analítica,
Nutrición y Bromatología
Facultad de Química empena@ull.es

ANN in Science: Entomology

J. Vaňhara, J. Havel & P. Fedor

1) Introduction

Identification of biota is fundamental in systematic biology and plays an important role in many practical fields, such as biological monitoring, agriculture, forestry, Nature conservation, medicine, etc.

For more than 250 years of modern taxonomy, the number of known species has been continually increasing, but our knowledge of biodiversity is far from being complete. There are vast arrays of morphologically similar species, which are often so small in dimensions as to render establishing the species identity within many groups a difficult task.

Identification usually requires significant experience, encyclopaedic knowledge, a good reference collection and relevant literature. The process of becoming expert in a group is mostly pain-staking and time-consuming and the numbers of available taxonomists are limited for many groups and zoogeographical regions.

To transform the taxonomic process it is necessary to increase the productivity of identification of biodiversity, including searching of new species by using new tools. Among such methods, the Artificial Neural Networks (ANN) methodology, as the artificial intelligence principle, has been used by us, as a more efficient tool based on the classical mostly morphological principles known from the time of Linnaeus.

In addition to use in taxonomy, ANN has also uses in ecology. It is possible to simulate biocoenosis structure under certain environmental conditions and also to define habitats by way of species.

2) Cooperation

From project *Tachina* were prepared 2 Bachelor, 2 Master and 1 Ph.D. Theses. Both Master Theses students were awarded by Dean´s Price in 2008. With all students as co-authors were published 3 tachinological papers based at least partly on ANN. It was about 5 years of the work.

In Thysanoptera project 2 Ph.D. students participated. Till now were published 2 papers.

During specialized zoological congresses 8 communications were presented in Australia, Brasil, Canada, Czech Republic, Japan, Slovenia, Slovak Republic,and South Africa.

ANN studies are part of the project of the Ministry of Education of the Czech Republic and Masaryk University Brno, MSM 0021622416

3) ANN in Entomology

Till now, the use of ANN has spread to many branches of Science but in Entomology or close Arachnology, as far as we know, is still rare. If used they are probably the most common and currently employed in insect taxonomy, but very often based on different and in many times no-traditional characters.

ANN offer a new innovation and rapid approach to routine identification of biota than traditionally dichotomic or pictorial keys and some statistical methods.

Much effort is required to create the training data set with relatively large number of specimens correctly identified by specialists. The more correctly identified specimens included in the training the better. This is because variation. On the other hand, if good diagnostic characters are selected, the ANN identification might be possible with fewer specimens of a particular species. However, once created this method show many advantages.

4) Use of ANN in Biology

Compared to traditional statistics, ANNs are non-linear and can describe highly complex systems such as found in biology. The system is based on a supervised ANN classifier.

From biological point of view, ANNs are analogues to biological neural networks and can simulate learning and solving of problems.

ANN system requires a good trained database, an essential prerequisite to obtain reliable identification, in which specimens are rigorously identified largely on the basis of good characters.

An ANN model is designed to find relationships between characters (*input*) and species (*output*). An advantage is that it combines an ability to learn from examples with generalization of patterns that have been previously observed.

Despite the clear advantages, the use of ANN for species identification has been relatively infrequent.

5) Characters

Different types of characters have been used for ANN in taxonomy of insects (and spiders), e.g. qualitative characters, colour, presence or absence of different body structures (see Thysanoptera), together with quantitative characters, e.g. meristic, morphometric (Thysanoptera, Diptera – Psychodidae, Tachinidae) or sexual characters (Thysanoptera,

Diptera – Culicidae, Tachinidae) as well as digital images (Lycosidae, Aphidae, Culicidae, Hymenoptera, and Orthoptera).

Methodology outside insect taxonomy have also used data that had been obtained by other methods, e.g. by spectrophotometry, climatology, chemometry, etc. In the framework of a particular insect group were analyzed sounds of Orthoptera. Aphid species (Aphidae) were tested by wing flaps by digital registration with an optic sensor. ichneumonids (Hymenoptera) were analyzed by digital images.

6) Entomologist starts with ANN using

ANN used for identification requires carefully built and trained database. Each specimen has to be characterized by a set of assorted characters that are adequate for the classification (not identification, see further slide) of the species, genus or other taxa.

Once such a database exists, a model is designed to find a relationship between characters (=input) and species (=output).

The next phase is called learning or training. During training, the outputs approximate the target values that are given as the inputs in the training set. Among various learning methods in ANN, computing the Back Propagation algorithm is the most popular.

The reliability of the database and the species identification must be checked using cross validation.

7) Database creation as a fundamental step

Database is built on identified specimens and of characters which were used for identification or any other, which are based on various types of data and couldn't have be related to identification.

Our ANN approach is based on measurements of morphological characters, structural shapes or colour. Quantitative characters were measured as linear distances on digital images taken from slide mounted and/or dry mounted specimens.

Morphometric access used by us is a non-destructive tool and is suitable also for type material and permanently preserved material.

Many species of Thysanoptera exhibit a pronounced sexual dimorphism. Males of some studied thrips are brachypterous or apterous and lack ocelli. This data were included into the data matrix as zero values. The same approach was chosen for the characters on ovipositor, which are applicable only to females. To distinguish between males and females, the sex was included as another two-state input character into the ANN insect model.

The form of database, or specialized sub-database selected from a main database, should be like this cut-example:

No	Character 1	2	3	4	5	6	7	8	9	10	sex	spp.
1	66.3000	284.7000	140.4000	62.4000	37.0500	25.3500	148.2000	Y	N	0	m	albi
2	66.3000	284.7000	140.4000	62.4000	39.0000	25.3500	148.2000	N	Y	0	m	albi
3	68.2500	280.8000	136.5000	64.3500	35.1000	27.3000	144.3000	N	Y	0	m	albi
4	68.2500	288.6000	138.4500	66.3000	35.1000	25.3500	150.1500	+	-	0	m	albi
5	70.2000	292.5000	138.4500	60.4500	37.0500	25.3500	150.1500	+	-	0	f	albi

etc.

8) Search for the optimal ANN architecture

ANN architecture should be conventionally the simplest, according the number of taxa. Multilayer perceptrons networks (MLP) is mostly constructed as three or four-layered (xx, n/nn, spp.), where *xx* is a number of characters in the input layer, *n/nn* are numbers of nodes in the one/two hidden layers, and *spp.* are the numbers of taxa in the output layer.

A smaller part of samples were randomly removed from the original data to compose the test set. These samples were reserved for a single unrepeated testing of the final ANN model, which had been selected and optimized using the learning and the verification sets. The test set did not play any role in the search of the model and ensured that the results of the training and verification sets were real and not artefacts of the training process.

The remaining part of samples was used to search for a suitable and optimal ANN architecture and for corresponding network training.

Optimal ANN architecture is independent on the number of samples in the training set. Even when 50% of the available data was used in the training set, the effect of *n* on the RMS error (see further) value remained very similar.

9) Number of nodes

The minimal value of Root Mean Square (RMS) was achieved according to the graph. As is usually recommended we chose the number of nodes *n* slightly higher (one or two nodes) than the optimum found. For further computations use only such *n*, where nearly or 100% correct classification of samples in the training process was achieved in most of the runs.

As said above, another suitable architecture found in training experiments could be with two hidden layers, which could also give good results and a 100% prediction. However, the simpler three-layered architecture was preferred by us, if it was found to be experimentally more resistant and robust in performance.

10) Training

SW Trajan uses and compares characters of every taxon with characters of all taxa at the very moment. Training of neural network should be successful if all our data are OK.

The training of a MLP network can be executed by different algorithms. We used the Back propagation, which is the best-known one and has relatively low memory requirements. We ran the training algorithm several times for 5,000 to 10,000 iterations (epochs).

11) Verification.

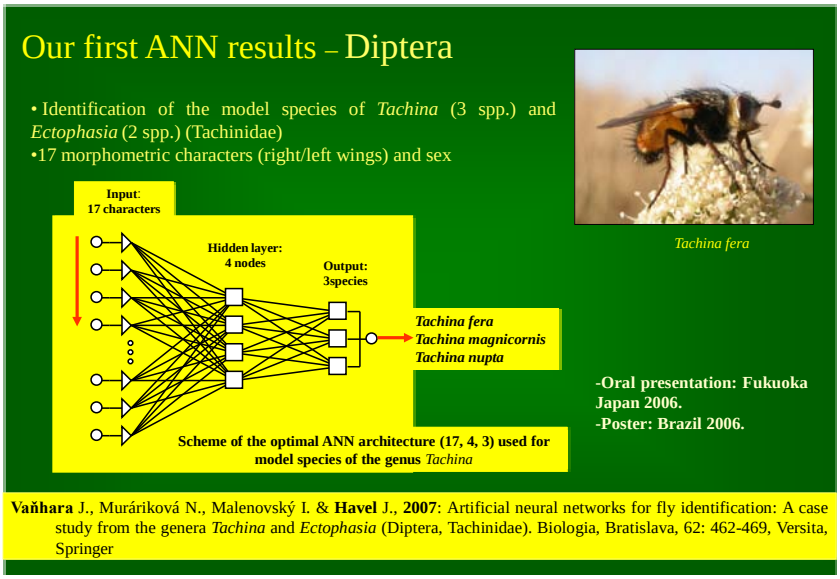
After obtaining the optimal architecture and minimal RMS, a number of randomly selected specimens from the learning set were excluded to form the verification set. They were used as a check for cross-validation of the training procedure to prevent over-training - the situation when the model is too complex and training achieves a low error but has a poor generalization when new samples are processed. The verification is a test of prediction power of the model; its efficiency is in identifying unknown specimens. However, unlike the test set, the verification set does actually play a role in selecting the final ANN model.

12) Identification

A properly trained and validated model can be applied for identification of unknown specimens and is able draw attention also to new species (sp.n.) as in our projects under study. Real description according International Code of Zoological Nomenclature needs specialized entomologist.

13) Own first ANN results: *Tachina, Ectophasia*

We have used ANNs as a tool for the identification in recent research projects on various insect groups. The first has included ANN methodology and was applied to the parasitic flies of the well known genera *Tachina* and *Ectophasia* (Tachinidae, Diptera). Two databases with 3 and 2 model species, 17characters, ANN identifications were fully successful.



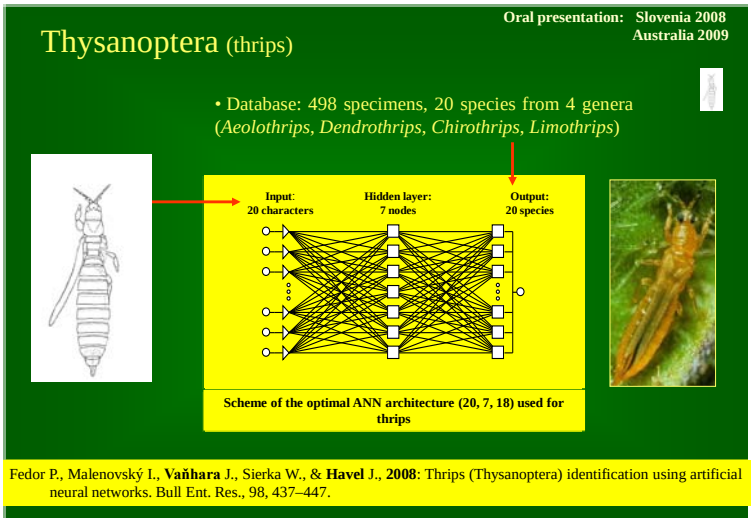
14) Thrips

The second entomological project undertaken by us has been on application of ANN for morphometric identification of 18 common European thrips species (Thysanoptera) from 4 genera. Thrips are minute insects, some of which are significant agricultural pests (e.g. vectors of harmful tospoviruses). Their correct identification is often needed to assess economic risks or to decide on control or quarantine measures. Similarly, as with the parasitic flies, there are relative few specialists who can deal with the taxonomy of thrips.

We used 20 input characters (different lengths measured on the thrips body, simple presence/absence of characters and sex). Three layer perceptron architecture achieved 97 % correct identification of 18 species, including two pests which attack cereals and may cause serious grain losses.

Our experiment on thrips suggests that identification is possible even if only a few specimens of a species are included into the training data set if a good measured data are enclosed (the lowest number of specimens was 9).

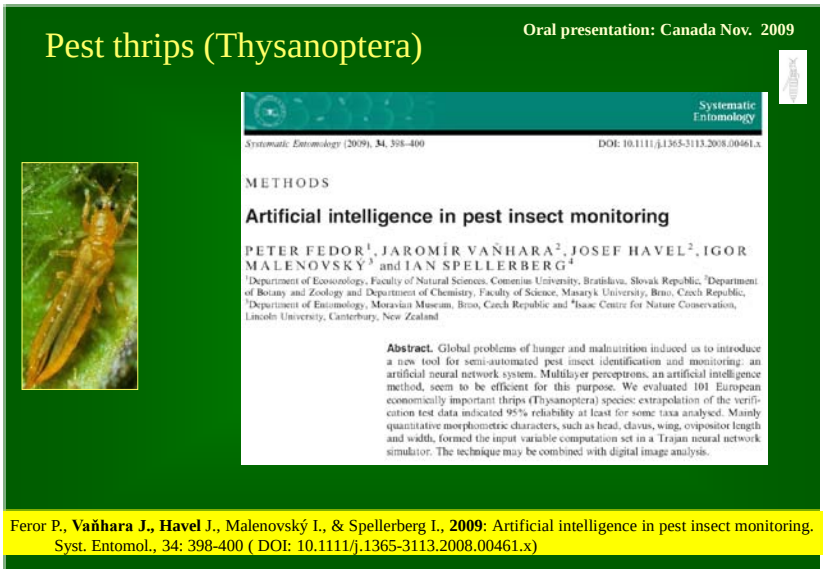
The same is true for the basic principle how to select and prepare data, how to select an appropriate network and how to interpret the results. The level of user knowledge needed for a successful application is however lower than in many other more traditional statistical methods and ANN are intuitively appealing.



Fedor P., Malenovský I., **Vaňhara** J., Sierka W., & **Havel** J., **2008**: Thrips (Thysanoptera) identification using artificial neural networks. Bull Ent. Res., 98, 437–447.

15) Pests, crop damage and ANN

Precise and prompt identification is essential for effective monitoring of phytopathogenic species. This includes the EPPO (European and Mediterranean Plant Protection Organization) quarantine pests that would otherwise cause crop damage. As shown in the thrips example, semi automated ANN data evaluation has offered a new tool for the easy and rapid monitoring of pests. A complete dataset of economically important tospovirus vectors such as *Thrips palmi* and *Frankliniella occidentalis* could be prepared as an advanced technological tool for pest monitoring. Recently, relevant morphometric and qualitative characters of European Thysanoptera pests are analysed. This will include quarantine species and will be used to construct suitable process for prompt and effective identification. This project will solve several practical identification problems in biosecurity and make identification more reliable. The system is suggested to be of considerable benefit in the Phytopathology applications.



Feror P., **Vaňhara J., Havel** J., Malenovský I., & Spellerberg I., **2009**: Artificial intelligence in pest insect monitoring. Syst. Entomol., 34: 398-400 (DOI: 10.1111/j.1365-3113.2008.00461.x)

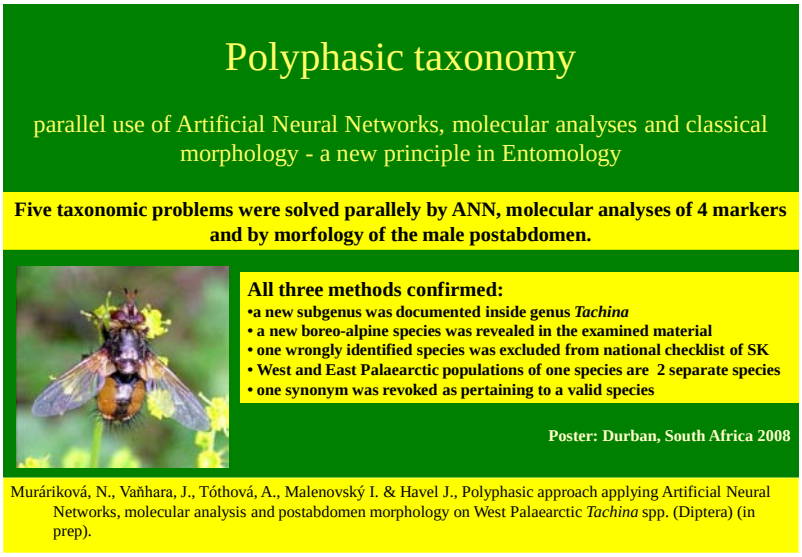
16) Polyphasic approach

The identification of the genus *Tachina* is generally considered difficult because diagnostic characters often overlap among species (colour, different length ratios, etc.). In spite of these difficulties, our *Tachina* project used ANN methodology successfully, not only for identification but also for solving several taxonomic problems, see slide.

The correct identification and conclusions on the systematics were based on ANN methodology, results were also tested by molecular analysis and by comparative morphological examinations incl. Cladistics. In all cases, the same conclusions were obtained.

Multidimensional combination of ANN and two alternative methods is homology to polyphasic taxonomy which is common in microbiology and in phylogenetics is termed total evidence. We believe that this is the first time that the accuracy of the ANNs methodology has been demonstrated against other more traditional methods. The principle of integrating different sources of data used for identification was initiated in 1970, when was introduced the term "Polyphasic Taxonomy" into Microbiology.

Such a polyphasic approach takes into account all known phenotypic and genotypic information and integrates them not only for the purpose of taxonomy and identification, but also for the full reciprocal validation of methods used and for the results obtained.



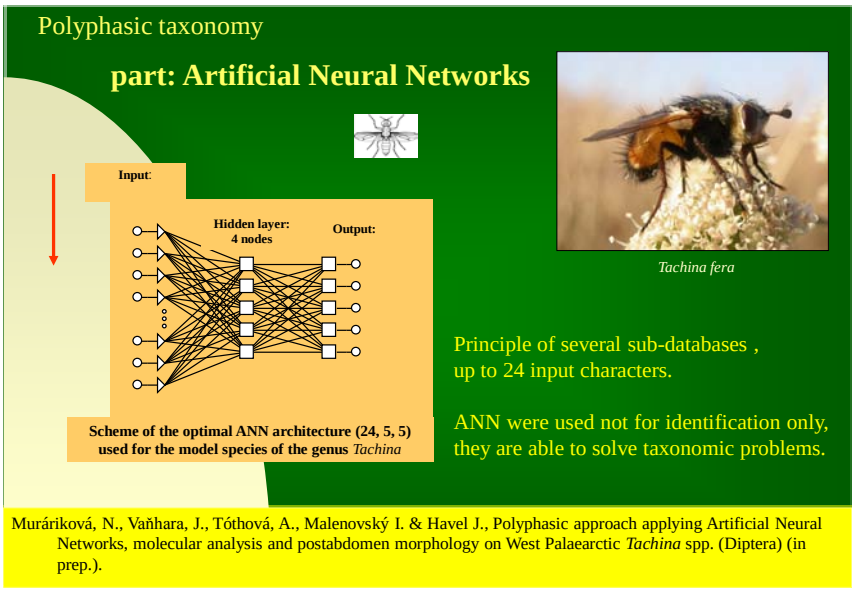
17) The role of ANN in polyphasic approach

Artificial neural networks (ANN) methodology, molecular analyses and comparative morphology of the male postabdomen were used successfully in parallel for species identification and resolution of some taxonomic problems. The supervised feed-forward

ANN model based on 24 input morphometric characters (selected on wing, antenna and male post-abdomen) led to a reliable identification.

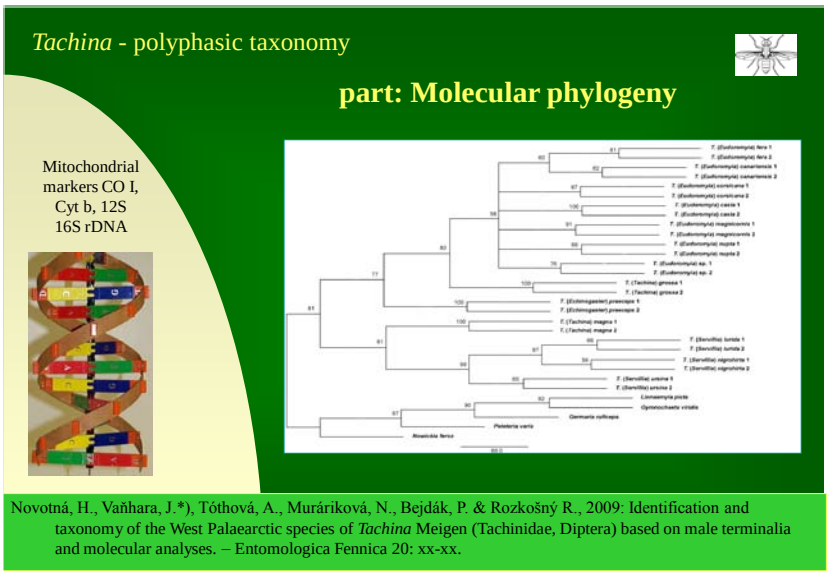
Procedure of Artificial Neural Network was based on the principle of 3 sub-databases based on Main database consists from about 500 lines.

Correctness of artificial neural network approach use in Taxonomy was justified by two independent methods (molecular and morphological) for the first time.



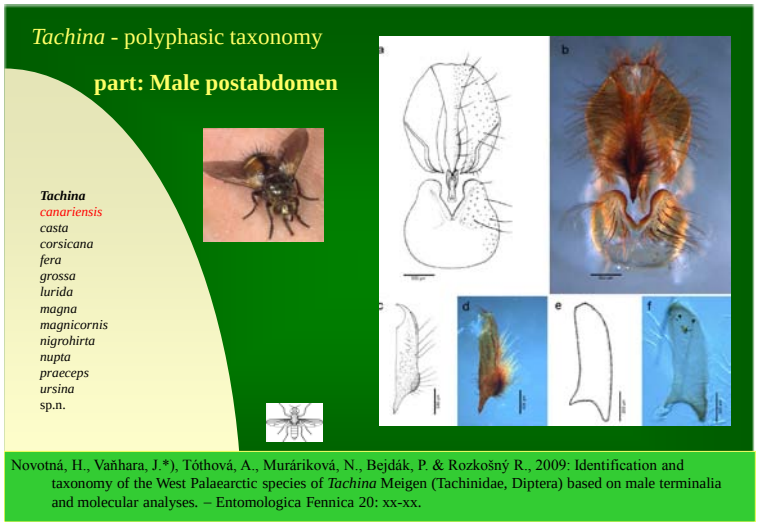
18) Molecular background

Molecular analyses were based on four mitochondrial markers CO I, Cyt b, 12S and 16S rDNA for subgeneric level and two markers 12S and 16S rDNA for the species analyses.



19) Morphological background

The knowledge of *Tachina* taxonomy and phylogenetic relationships of its species is still insufficient. West Palaearctic species of the genus comprises 12 valid species and also 45 synonymic names. That is why a new identification key was done, based on male postabdominal structures. All characters are illustrated by original pen drawings and deep focus micrographs, some of them. (*T. canariensis*, *T. casta* and *T. corsicana)* for the first time. Cladistics based on male postabdominal characters is consistent with molecular conclusions and ANN results.



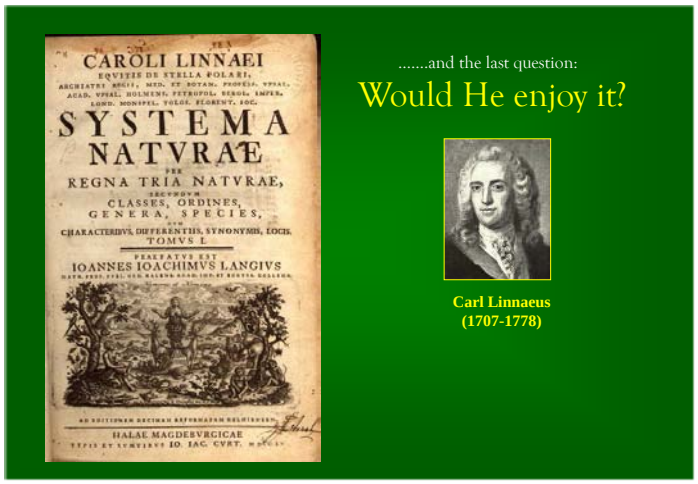
20) Conclusions

- ANN represents a new powerful tool for fast and (semi-)automated identification of insects (not only) and open great possibilities for taxonomy.
- ANN methodology can be applied for any entomological objects (not only) for which it is possible to determine characters (=variables) with sufficient information content to enable their resolution and correct identification .
- Identification of species is possible across different genera.
- ANNs are able to determine even hardly distinguishable species or to indicate new species.
- ANNs are able to solve taxonomical problems.
- A higher number of specimens for the training of ANN improves the quality of identification.
- It is better to use a principle of sub-databases instead of one big one.
- Sex has some effect on the input parameters and should be given as input.

- ANNs are able to take into account all the input characters simultaneously (similarly as multivariate statistics).
- ANN can replace some destructive analytic methods if the material must not be destroyed (e.g. types, material mounted on slides).
- ANN enable identification of numerous and uniform samples (e.g. insect pests) by technical staff, without the risk that some aberrant specimens or new taxa would be overlooked (ANN can indicate them) .
- Specialists (experts) thus could be saved for really expert tasks.
- Rare taxon specialists are often alone in their studies and they need the next independent opinion which ANN is able to provide.

21) Future

The use of ANN in taxonomy has some limitations. They require a sufficiently high number of specimens to construct an effective database. Further research is also needed for better understanding of the character selection and its applications. This research and assessment in these areas will almost certainly increase the popularity of the use of non-traditional methods for species identification.



22) References

Fedor P., Malenovský I., Vaňhara J., Sierka W. & Havel J. (2008): Thrips (Thysanoptera) identification using artificial neural networks. *Bulletin of Entomological Research* 98: 437-447.

Fedor P., Vaňhara J., Havel J., Malenovský I., & Spellerberg I. (2009): Artificial intelligence in pest insect monitoring. *Systematic Entomology* 34: 398-400. (DOI: 10.1111/j.1365-3113.2008.00461.x).

Muráriková, N., Vaňhara, J., Tóthová, A., Malenovský I. & Havel J., Polyphasic approach applying Artificial Neural Networks, molecular analysis and postabdomen morphology on West Palaearctic *Tachina* spp. (Diptera) (in prep).

Vaňhara J., Muráriková N., Malenovský I. & Havel J. (2007): Artificial neural networks for insect identification. *In: O'Hara J. (Ed.), The Tachinid Times* 20: 8-9.

Vaňhara J., Muráriková N., Malenovský I. & Havel J. (2007): Artificial neural networks for fly identification: A case study from the genera *Tachina* and *Ectophasia* (Diptera, Tachinidae). *Biologia* 62: 462-469.

This Manual was prepared within the Erasmus Programme, the exchange between University of La Laguna, Tenerife, Spain and Masaryk University in Brno, Czech Republic, TEACHING PROGRAMME 2009/2010.