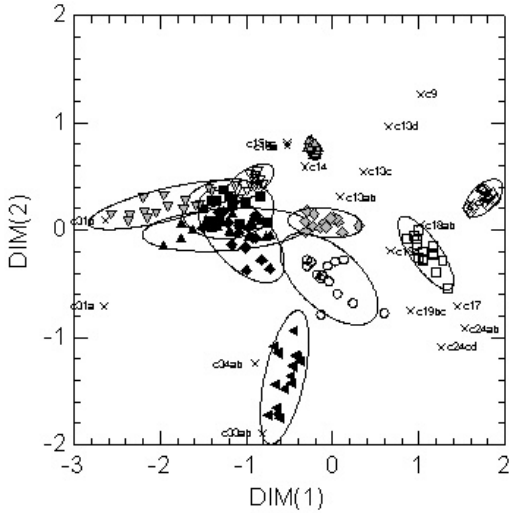
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION

6. ARTIFICIAL NEURAL NETWORKS

Prof. Eugenio Parente
Scuola di Scienze Agrarie - Università della Basilicata



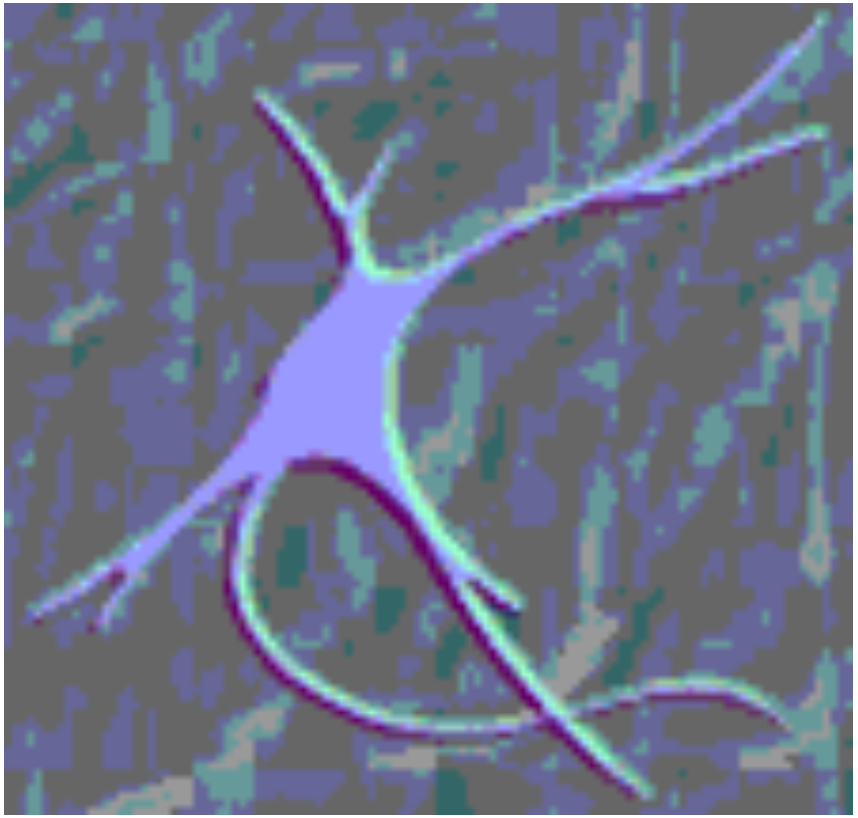
Outline

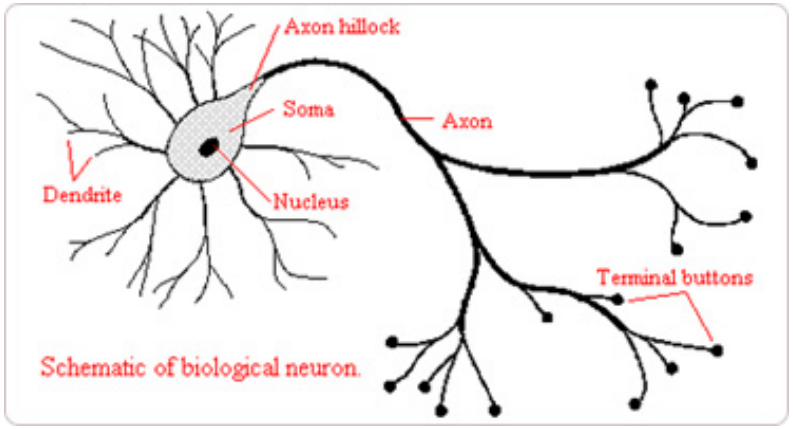
- definitions
- artificial neurons
- unsupervised artifical neural networks (Kohonen networks)
- supervised artifical neural networks (MLP, RBF) for regression and pattern classification

Definition

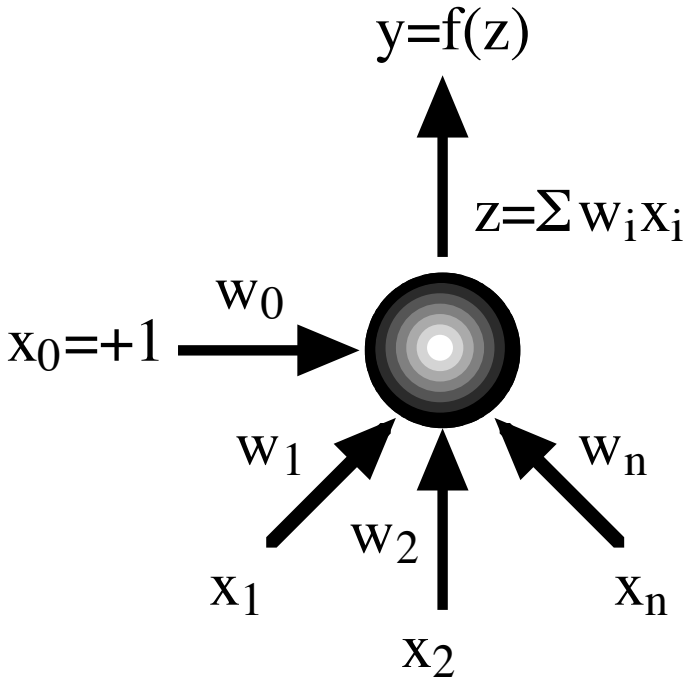
“a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects: (1) knowledge is acquired by the network from its environment through a learning process. (2) Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge” (Haykin, 1999)

Neurons



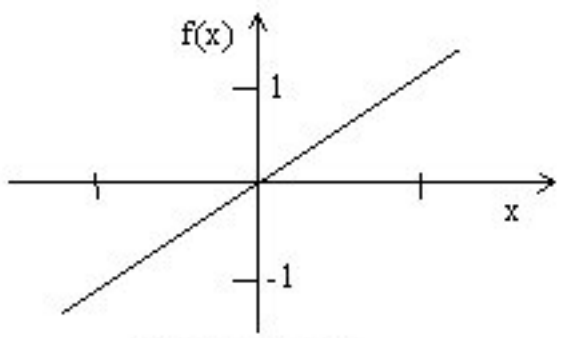

Axon hillock
Soma
Axon
Dendrite
Nucleus
Terminal buttons
Schematic of biological neuron.

Artificial neurons

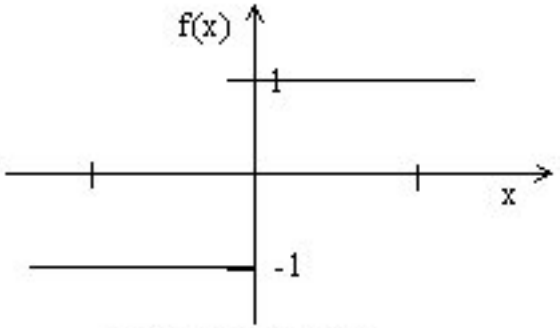


x_i	neuron inputs
w_i	synaptic weights
y	neuron output
$f(z)$	activation function

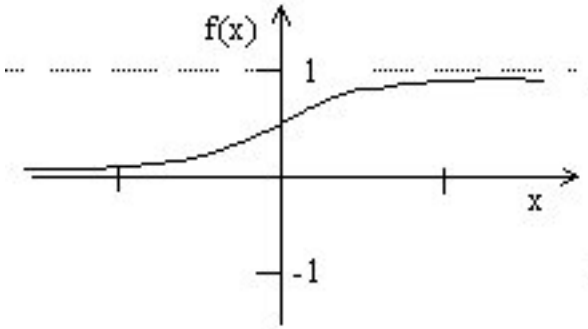
Transfer functions



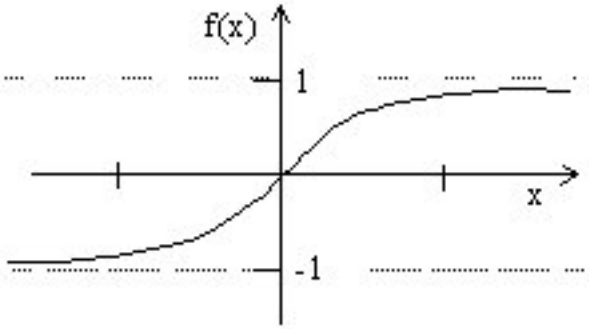
Linear: $f(x) = x$



Step: $f(x) = sgn(x)$

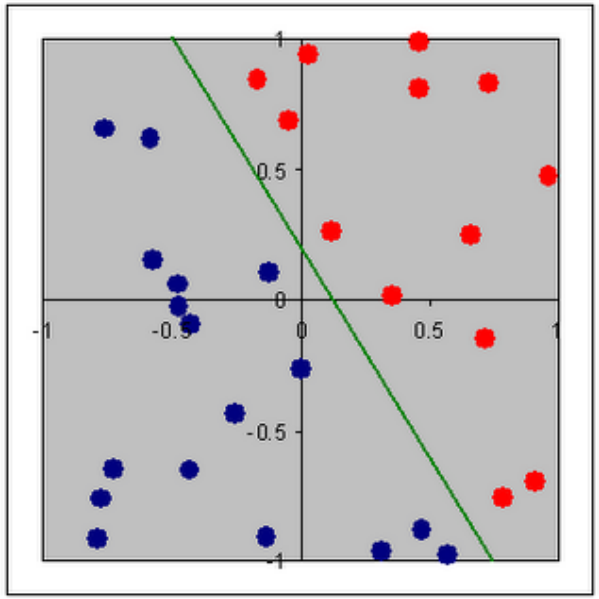


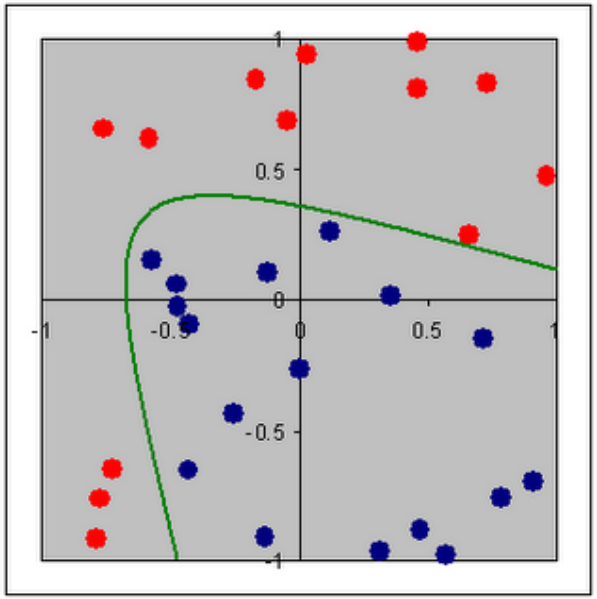
Sigmoid: $f(x) = 1/(1+exp(x))$



Tan-hyperbolic: $f(x) = tanh(x)$

Linearly separable and non-linearly separable problems





Supervised vs. unsupervised artificial neural networks

- **Unsupervised artificial neural networks**: the network is presented with the inputs during the training stage but it is allow to build its own representation of the data
- **Supervised artificial neural networks**: during the training stage each input is paired with the "correct" answer and the weights in the network are adjusted in order to minimize some sort of error function

Unsupervised vs. supervised artificial neural networks

- Unsupervised artificial neural networks can be used for data partitioning and unsupervised pattern recognition; they are of limited used for predictions
- Supervised artificial neural networks can be used for:
  - Supervised pattern recognition (symbolic output; continuous, discrete and/or symbolic inputs)
  - Regression, prediction: continuous, or discrete quantitative inputs and outputs
  - Time series analysis, backcasting, forecasting

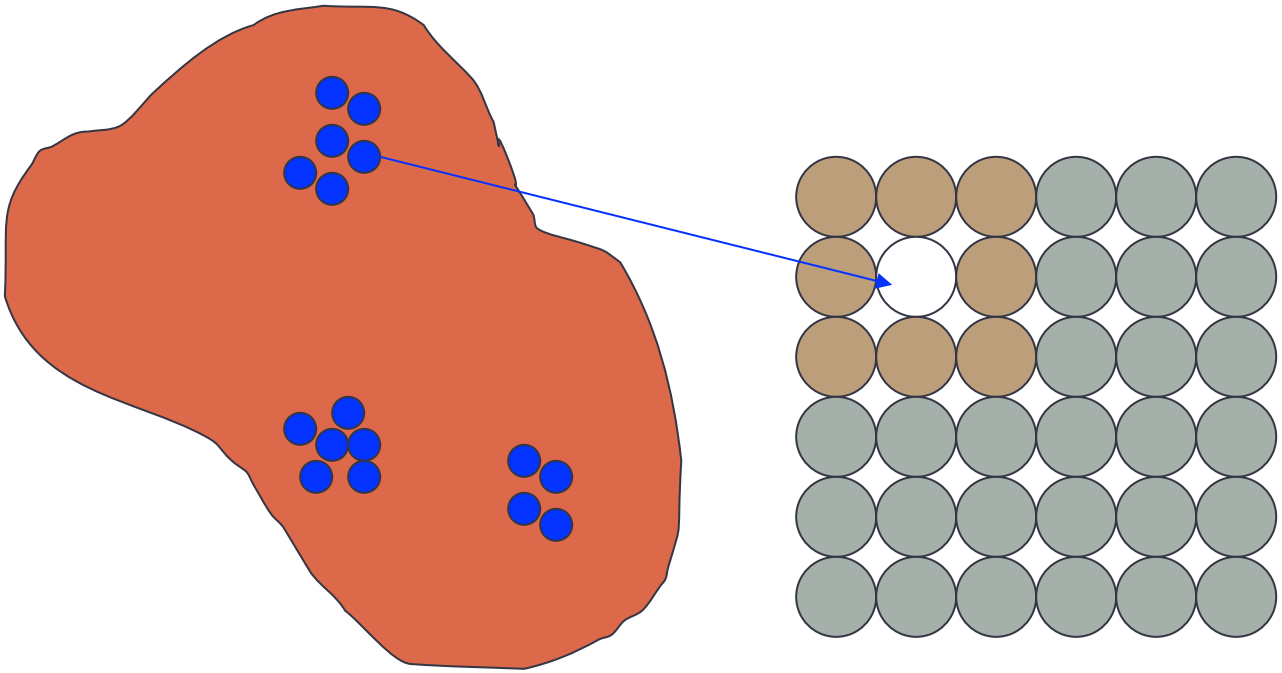
Important properties of properly trained ANNs

- ability to generalize, i.e., to provide reasonable outputs to inputs not seen before;
- ability to process nonlinear problems, due to the presence of multiple layers of neurons and/or to the use of nonlinear activation functions;
- fault tolerance, i.e., ability to produce reasonable outputs even if inputs are degraded (for example, because of missing or inconsistent data)

Supervised training by backpropagation

- The data set is divided in three groups (usu. 80:10:10): training, validation, test set
- The training set is used for training, the validation set to avoid overfitting and loss of generalization the test set to validate the results
- The network is initialized with small random weights and presented witht the test set inputs coupled to the desired outputs
- An error measure is calculated between network output and desired outputs and the weights of the layers (starting from the last, backward) are adjusted by some gradient descent technique to reduce the error
- The procedure is repeated until a convergence is obtained (no further change in error beyond a tolerance factor); the results on the validation set are also calculated to stop training when error in the validation set starts to increase
- The network is evaluated on the basis of results on the test set (crosstabulation, calculation of MSE)

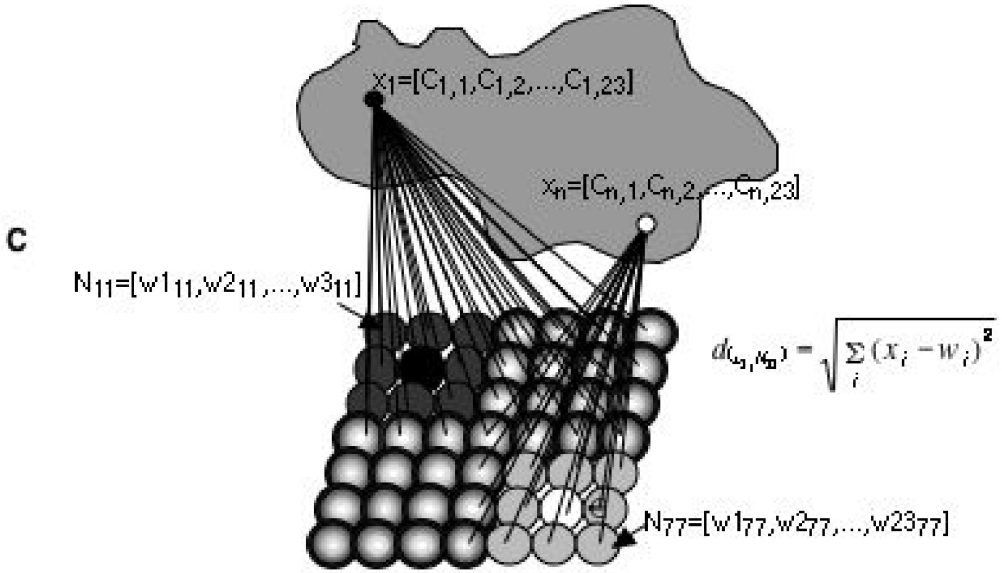
Kohonen self organizing map



input layer:
$\mathbf{x}_i$ (1<i<n)

output layer:
k neurons (in a square grid) each with a *p* dimentional weight vector **w**

Kohonen self-organizing maps



ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT

Journal of Microbiological Methods 66 (2006) 336 – 346

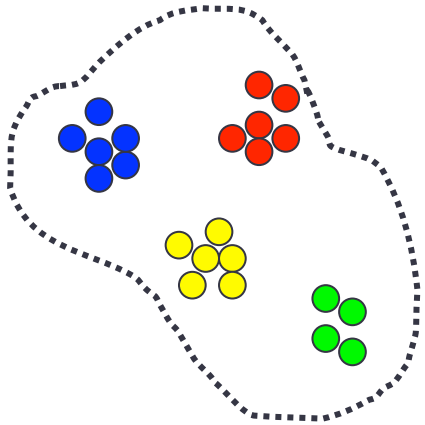
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

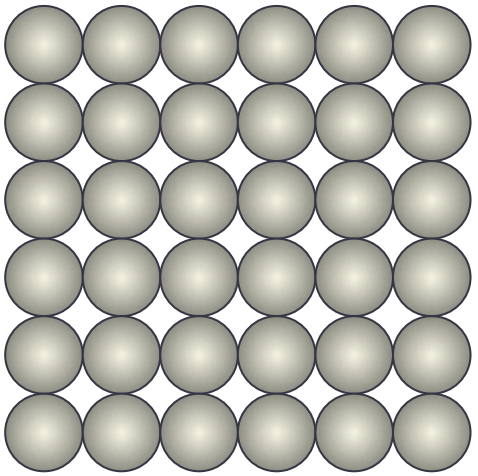
Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins

P. Piraino, A. Ricciardi, G. Salzano, T. Zotta, E. Parente *

Training of Kohonen self organizing maps (1)

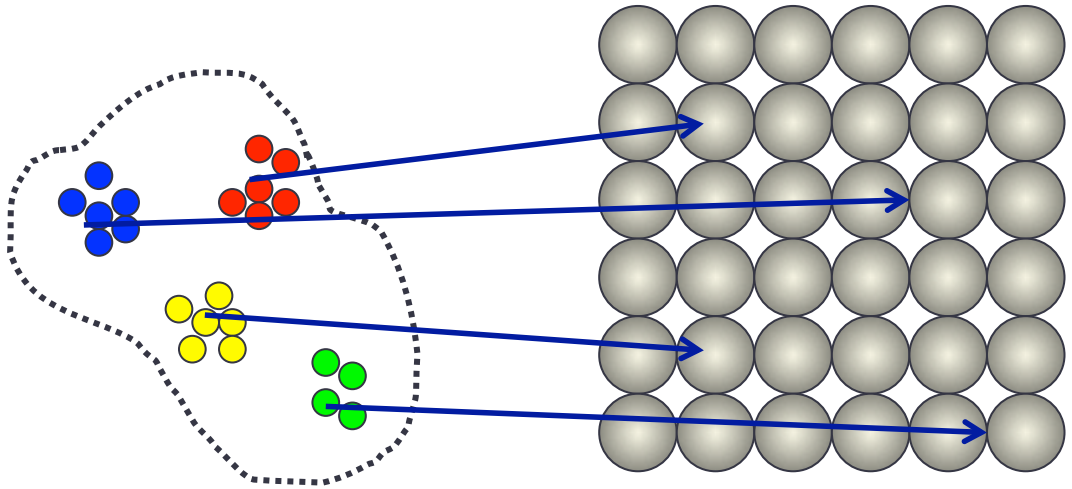
Clusters of *n* objects in a *p*-dimensional space; the position of each object i ($1 \leq i \leq n$) is defined by the vector $\mathbf{x}_i$ containing the standardized values for the *p* variables for which (continuous) measurements have been taken





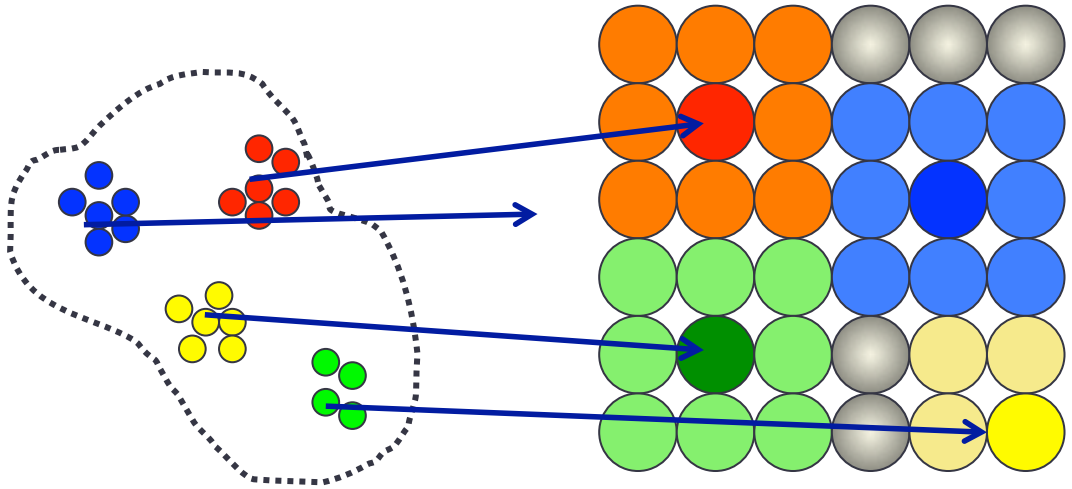
A square grid of *k* neurons each with a *p*-dimensional weight vector **w**, which is initialized with random numbers

Training of Kohonen self organizing maps (1)



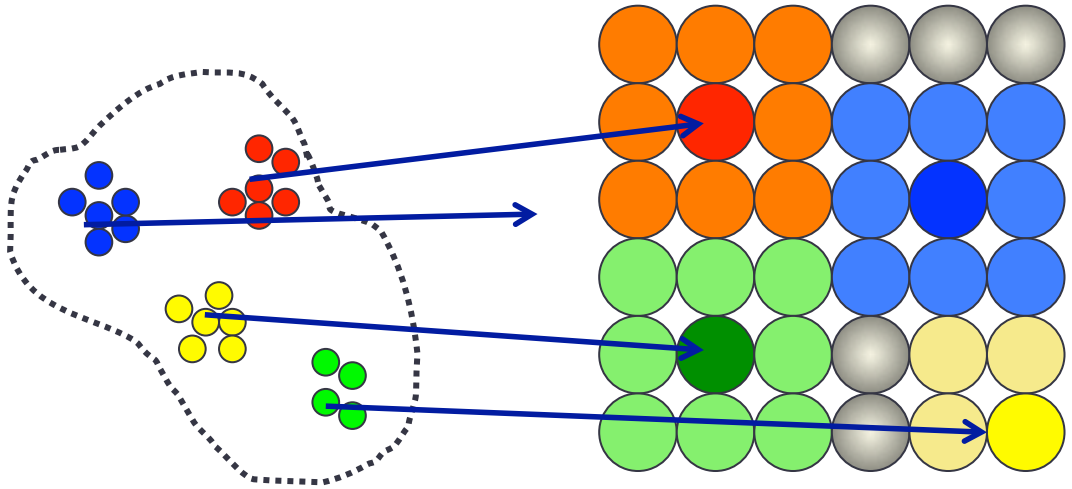
For each object x_i and each neuron y_j a distance measure is calculated between $\mathbf{x}_i$' and $\mathbf{w}_j$'

Training of Kohonen self organizing maps (1)



The weight vector $\mathbf{w}_j$' of the neuron which is closest to object $\mathbf{x}_i$' (the "winning" neuron) is updated to make it closer to x_i. The weights of neiboughring vectors are also updated.

Training of Kohonen self organizing maps (1)

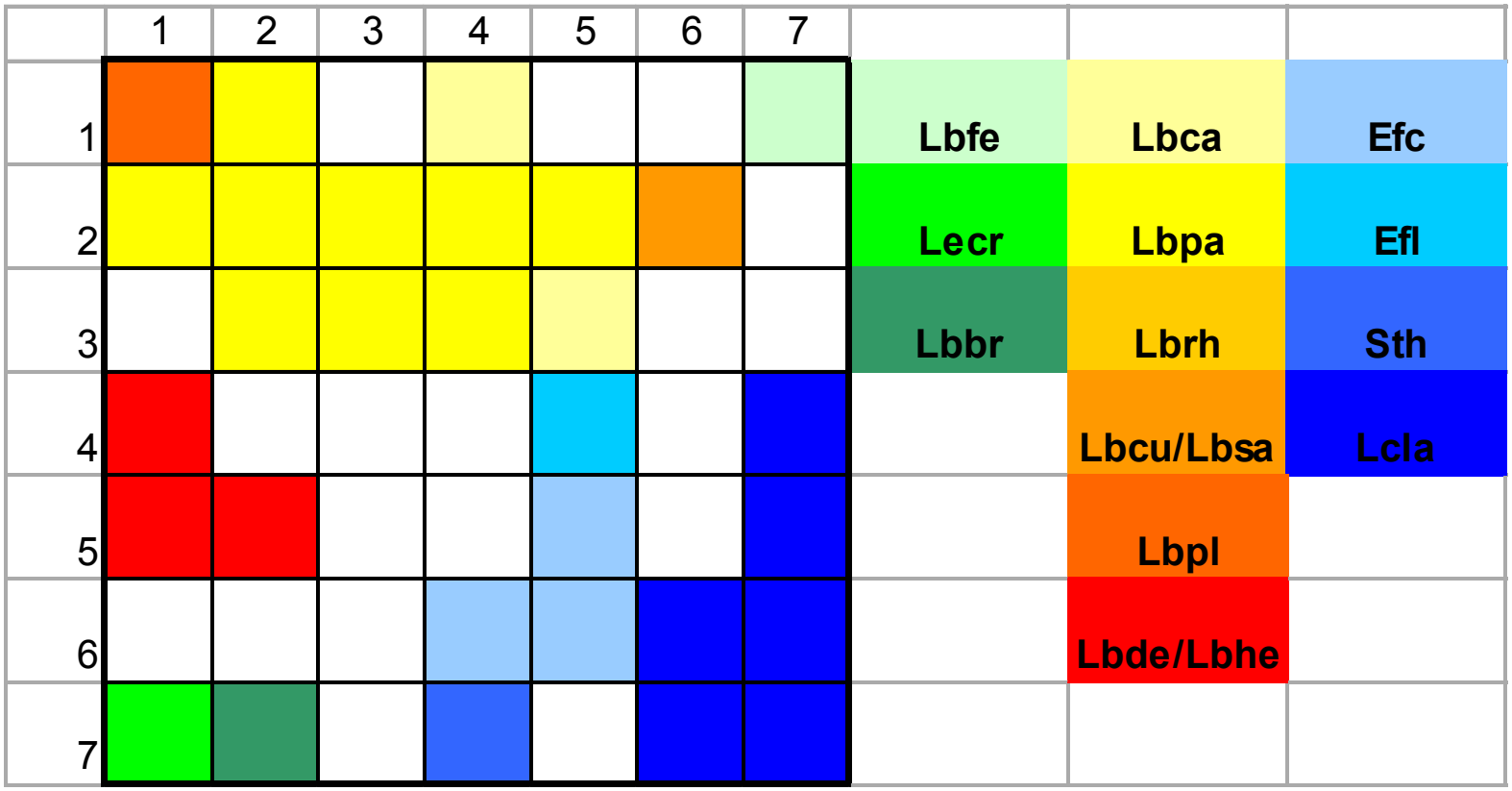


The process is repeated iteratively until convergence is obtained and no further change of $\mathbf{w}_j'$ is necessary. Each of the *n* objects maps (is closest to) one of the *k* neurons. Neurons or groups of neighbouring neurons represent clusters of objects.

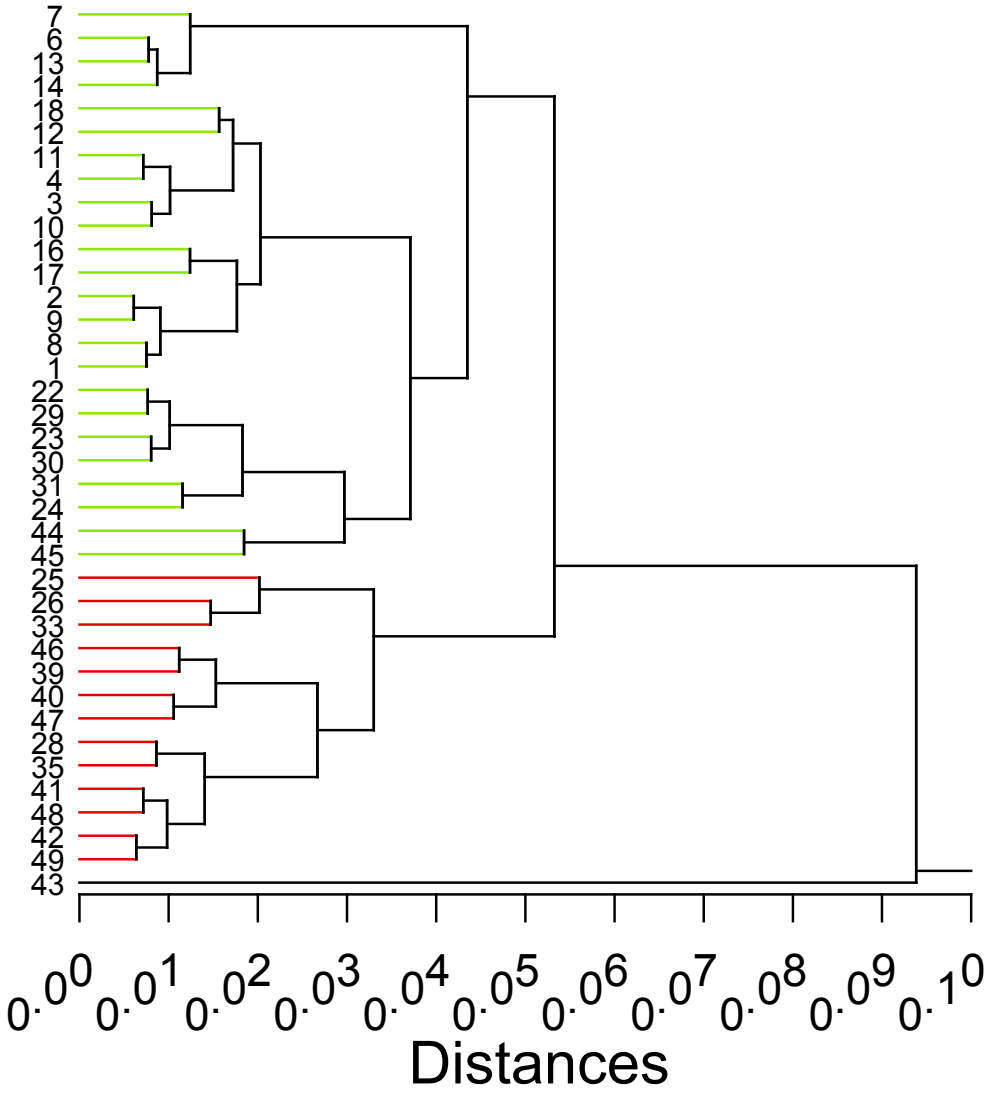
Useful properties of Kohonen networks

1. Kohonen SOMs are built in analogy with the organization of some areas of the brain which process external stimuli, and in which neurons responding to the same stimulus are close
2. after training the neurons are placed in the input space and mark clusters of data
3. Kohonen SOMs have some analogies with MDS and k-means but can accept a large variety of data
4. Kohonen SOMs can process very large amounts of data
5. Kohonen SOMs can be used with symbolic outputs to produce multilayered maps (one layer for each symbolic output)
6. in run mode Kohonen SOMs can be used to identify the node which responds more strongly to a new input

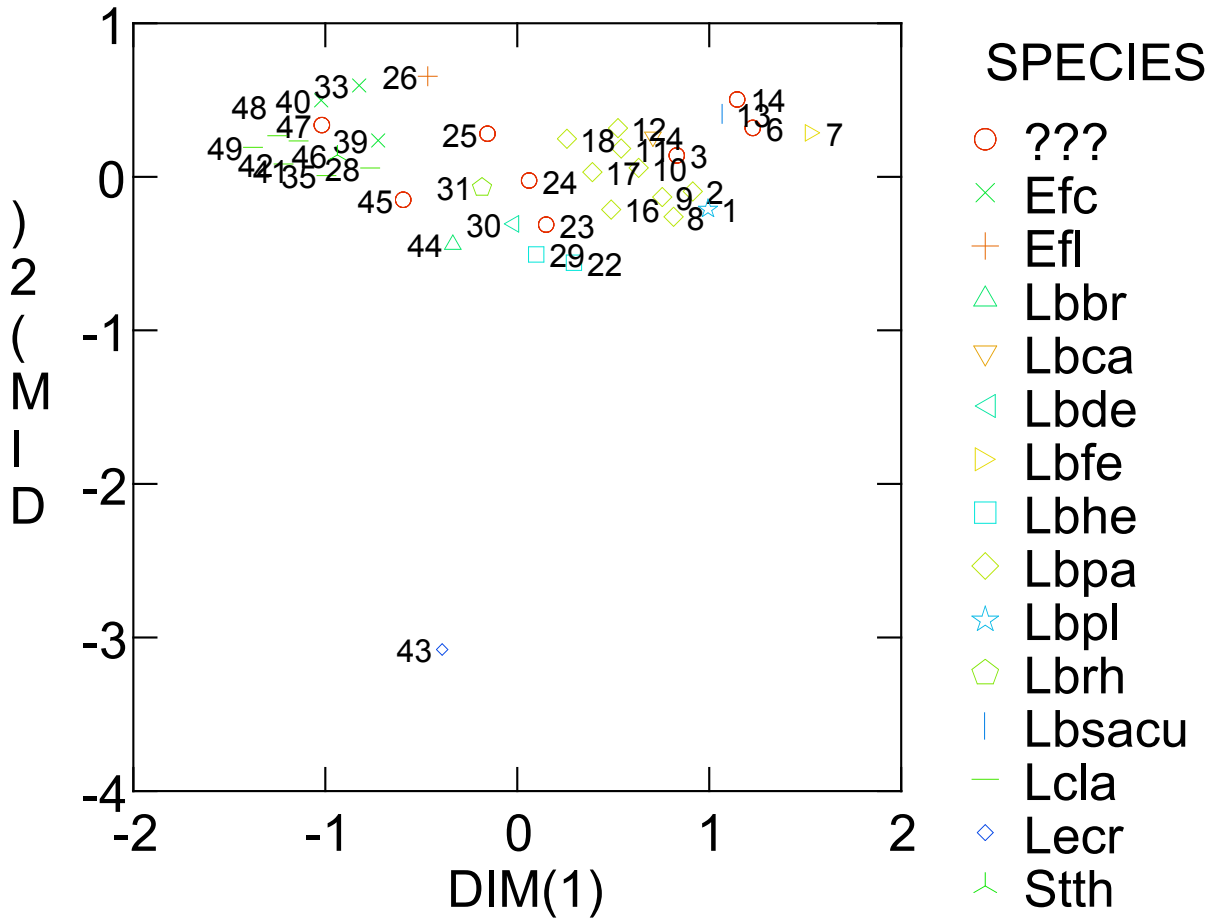
A Kohonen network for the classification of LAB on the basis of SDS-PAGE of WCP



Where are the nodes?



Where are the nodes?



1 Lb. brevis (12)	*2 Lb. delbrueckii ssp. lactis (10)*	*3 Lb. delbrueckii ssp. bulgaricus (10)*		*5 Lb. sakei (3)*		*7 Lb. fermentum (4)*
	9 Lb. rhamnosus (11)	*10 Lb. pentosus (2)*		*12 Lb. curvatus (3)*		*14 Lb. fermentum (4)*
		17 Lb. paracasei (5a)	*18 Lb. casei (6)*			
22 Leuconostoc spp. (13)					*27 Lc. raffinolactis Lc. lactis (15a)*	*28 Lc. lactis (15a)*
29 Leuconostoc spp. (13)		*31 Lb. paracasei (5b)*		*33 Ec. faecium (16d)*	*34 S. thermophilus (17b)*	
36 Leuconostoc spp. (13)				*40 Ec. faecalis + DPC1146 (16b)*		
	44 Lb. helveticus (8)	*45 Lb. helveticus (8)*				*49 Lb. plantarum Lb. paraplantarum (7)*

A Kohonen network for the classification of LAB on the basis of SDS-PAGE of WCP

ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Microbiological Methods 66 (2006) 336–346

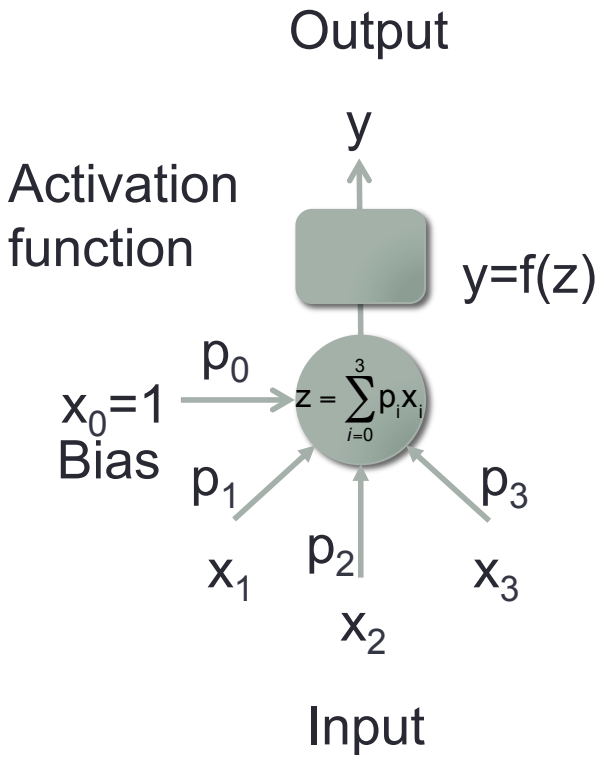
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins

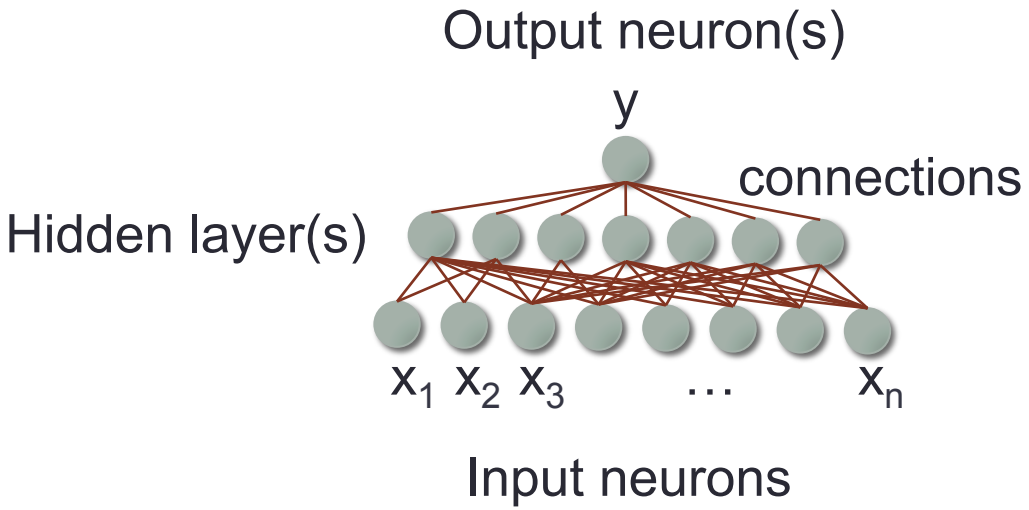
P. Piraino, A. Ricciardi, G. Salzano, T. Zotta, E. Parente *

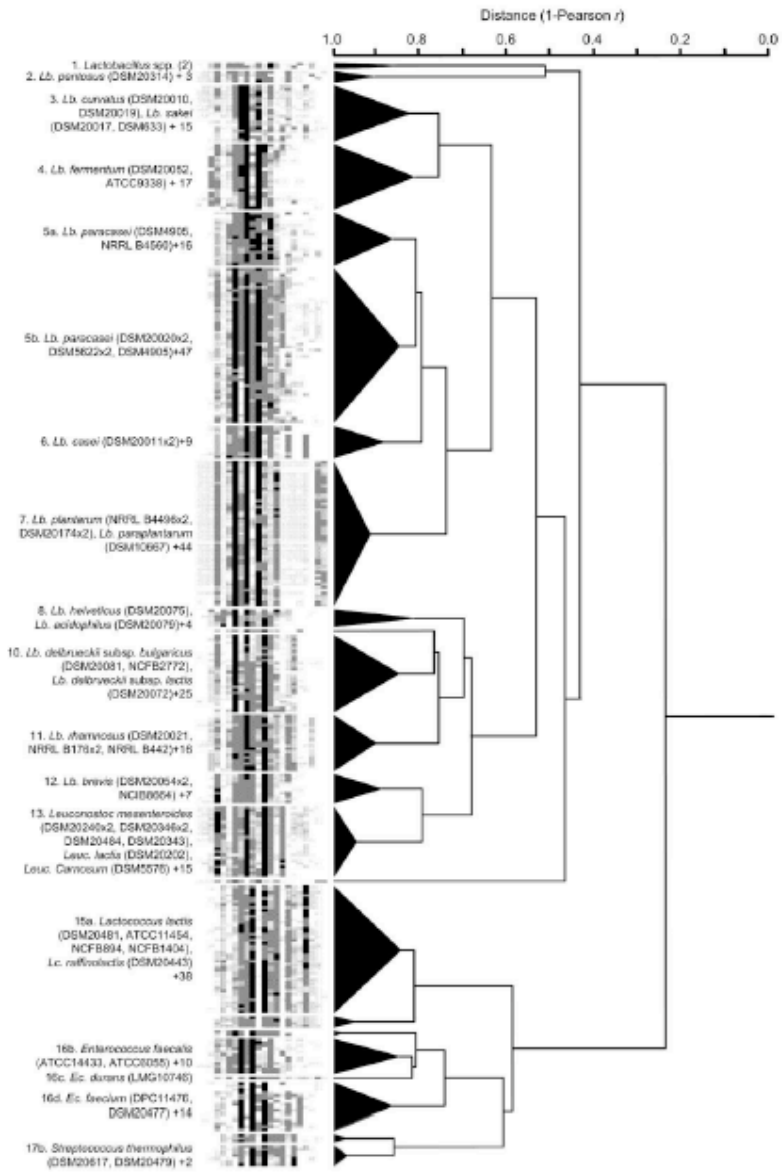
A multilayer perceptron

A perceptron

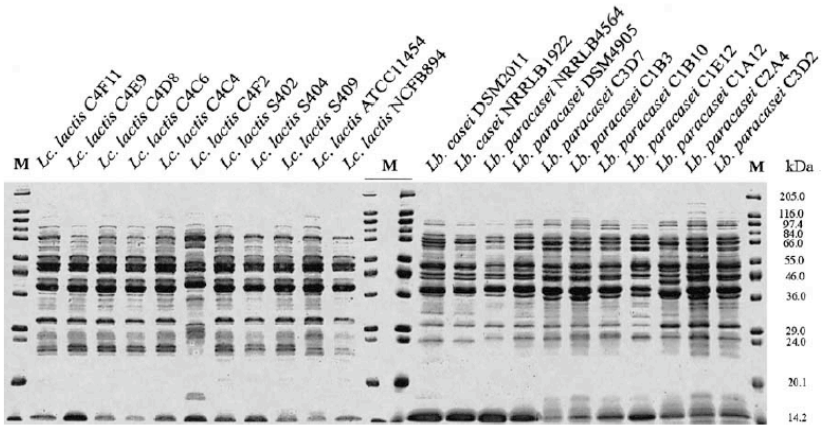


An artificial neural network
(amultilayer perceptron, MLP)





A supervised artificial neural network for the discrimination of whole-cell protein patterns



ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Microbiological Methods 66 (2006) 336 – 346

Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

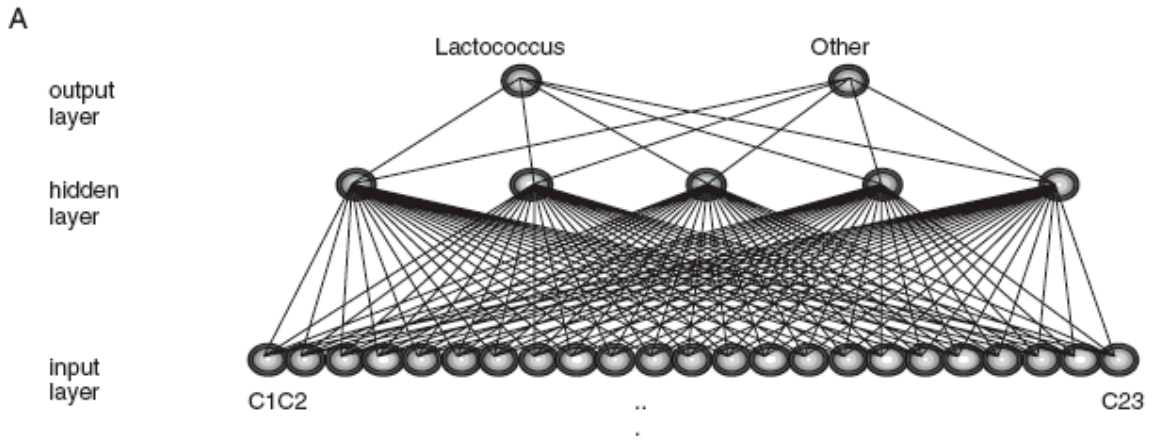
Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins

P. Piraino, A. Ricciardi, G. Salzano, T. Zotta, E. Parente *

Dipartimento di Biologia, Difesa e Biotecnologie Agro-Forestali, Università della Basilicata, Viale dell'Ateneo Lucano, 10, 85100 Potenza, Italy

Received 27 October 2005; received in revised form 16 December 2005; accepted 21 December 2005
Available online 15 February 2006

A supervised artificial neural network for the discrimination of whole-cell protein patterns



ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Microbiological Methods 66 (2006) 336 – 346
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins

P. Piraino, A. Ricciardi, G. Salzano, T. Zotta, E. Parente *

Dipartimento di Biologia, Difesa e Biotecnologie Agro-Forestali, Università della Basilicata, Viale dell'Ateneo Lucano, 10, 85100 Potenza, Italy

Received 27 October 2005; received in revised form 16 December 2005; accepted 21 December 2005
Available online 15 February 2006

A supervised artificial neural network for the discrimination of whole-cell protein patterns

Table 1
Percentage matching identifications, hierarchical cluster analysis (see Fig. 2) and Kohonen network, and % correct identifications for the test set for a Bayesian network and linear discriminant analysis trained to distinguish *Lactococcus* from other species

Species	% Matching, criterion 1	% Matching, criterion 2	% Correct, test set, 23:5:2 Bayesian network [a]	% Correct, test set, linear discriminant analysis [a]
Lb. brevis	80.0	100.0	100.0	100.0
Lb. casei	100.0	100.0	100.0	100.0
Lb. delbrueckii	92.3	92.3	100.0	100.0
Ec. faecalis	100.0	100.0	100.0	100.0
Ec. faecium	68.8	68.8	95.0	100.0
Ec. durans	n.a.	n.a.	100.0	0.0
Lb. fermentum	100.0	100.0	100.0	100.0
Lb. helveticus	100.0	100.0	100.0	100.0
Lc. lactis	87.2	91.5	100.0	100.0
Leuconostoc spp.	83.3	83.3	100.0	100.0
Lb. paracasei	42.9	71.4	100.0	100.0
Lb. plantarum	95.9	95.9	100.0	100.0
Lb. rhamnosus	52.6	52.6	97.5	100.0
Lb. sakei/curvatus	52.6	89.5	100.0	100.0
S. thermophilus	100.0	100.0	75.0	100.0
Lb. pentosus	50.0	50.0	90.0	100.0
Total	76.5	85.5	99.1	99.7

For the Kohonen network: with criterion 1 a matching identification is scored when an unknown strain activates the same neuron as the reference strain(s) belonging to the same cluster. With criterion 2 a matching identification is scored even if the unknown strain activates an empty neuron, provided that the empty neuron is closer to the neuron activated by the reference strain(s) with which the unknown shared a cluster in Fig. 2. Results are shown by species or group of species and for all strains.
n.a., not applicable.
[a] Average of 10 replicates.

ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Microbiological Methods 66 (2006) 336 – 346
Journal of Microbiological Methods
www.elsevier.com/locate/jmicmeth

Use of unsupervised and supervised artificial neural networks for the identification of lactic acid bacteria on the basis of SDS-PAGE patterns of whole cell proteins

P. Piraino, A. Ricciardi, G. Salzano, T. Zotta, E. Parente *
Dipartimento di Biologia, Difesa e Biotecnologie Agro-Forestali, Università della Basilicata, Viale dell'Ateneo Lucano, 10, 85100 Potenza, Italy

Received 27 October 2005; received in revised form 16 December 2005; accepted 21 December 2005
Available online 15 February 2006

A supervised artificial neural network for the discrimination of RAPD-PCR patterns

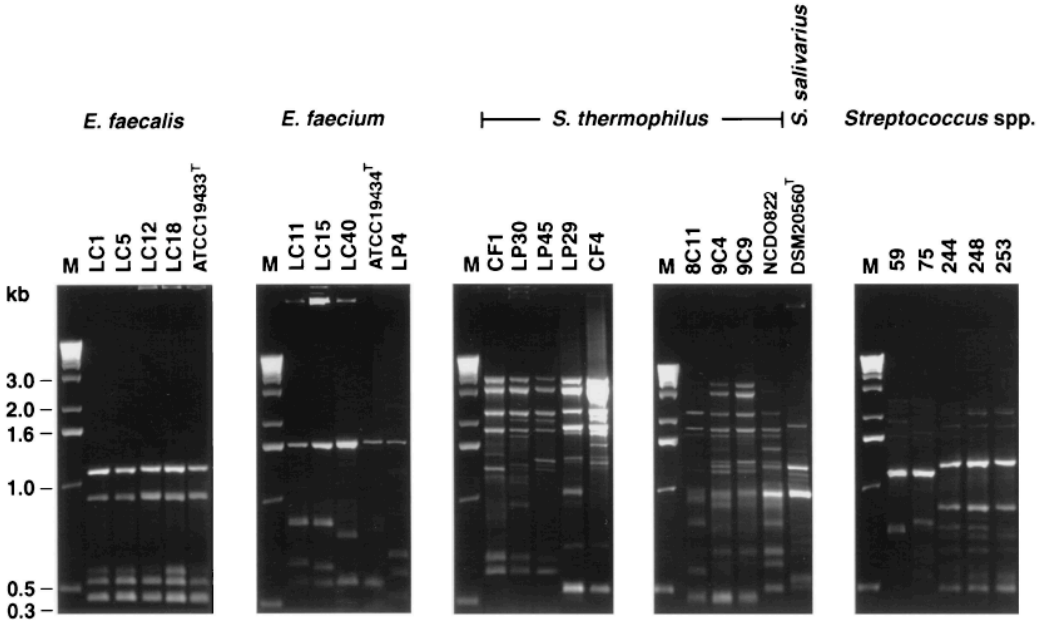


FIG. 2. Ethidium bromide-stained 1.5% (wt/vol) agarose gel displaying RAPD patterns of 32 strains of thermophilic streptococci obtained with primer XD9 (5′GAAGTCGTCC). Strain designations are shown above the lanes. Lane M, 1-kb DNA ladder (Gibco BRL) used as molecular size marker.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2001, p. 2156–2166 Vol. 67, No. 5
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.5.2156–2166.2001

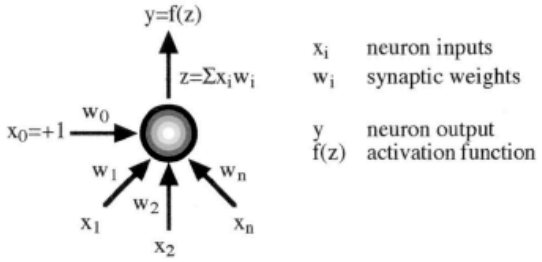
Comparison of Statistical Methods for Identification of *Streptococcus thermophilus, Enterococcus faecalis,* and *Enterococcus faecium* from Randomly Amplified Polymorphic DNA Patterns

GIANCARLO MOSCHETTI,[1] GIUSEPPE BLAIOTTA,[1] FRANCESCO VILLANI,[1] SALVATORE COPPOLA,[1] AND EUGENIO PARENTE[2]*

Dipartimento di Scienza degli Alimenti, Università degli Studi di Napoli "Federico II," 80055 Portici,[1] and Dipartimento di Biologia, Difesa, e Biotecnologie Agro-Forestali, Università degli Studi della Basilicata, 85100 Potenza,[2] Italy

A supervised artificial neural network for the discrimination of RAPD-PCR patterns

A. An artificial neuron



B. Architecture of Artificial Neural Networks used for the identification of thermophilic streptococci.

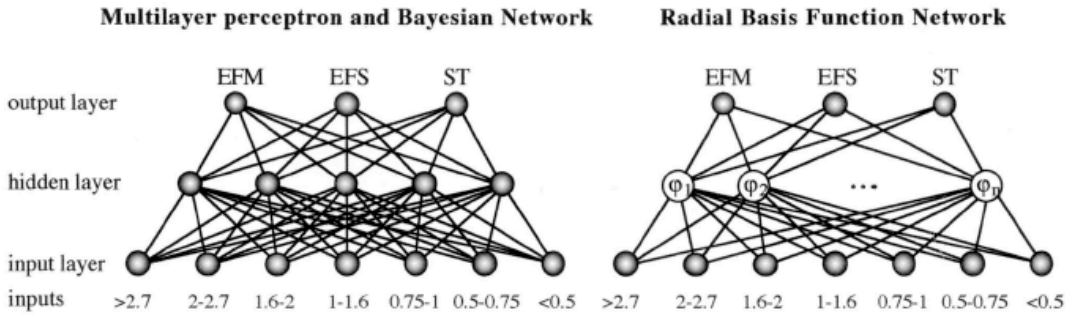


FIG. 1. (A) Schematic representation of an artificial neuron. The neuron is a simple processing unit connected to other neurons by synapses. A synaptic weight (w_i) is associated with each synapsis. An output y is produced by using the weighted sum ($z = \Sigma x_i w_i$) of its inputs (x_i; x_0 is fixed, and the product x_0w_0 is known as bias) as an argument of the activation function f(z). Different types of activation functions (nonlinear sigmoid functions as the logistic and hyperbolic tangent, but also threshold or linear functions) can be used. (B) Architecture of the ANNs used in this study. All types of networks used as an input the number of bands in selected molecular weight (in kilobases) intervals of the RAPD-PCR patterns and had three output nodes, one for each of the three species to be identified (EFM, *E. faecium*; EFS, *E. faecalis*; and ST, *S. thermophilus*). Both the MLP and the BN had a hidden layer with five nodes and used hyperbolic tangent activation functions, but they differed in the algorithm used to iteratively adjust the synaptic weights during supervised training (see the text for details). The hidden layer of the RBF was made up of 25 centers. For each of these, the Euclidean distance between an input pattern and the center was used as an argument of a nonlinear radial basis function, and the result was passed to the output nodes, which in turn had a linear activation function. The number and coordinates of the centers in the input space and the synaptic weights of the output neurons were adjusted during supervised training.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2001, p. 2156–2166 Vol. 67, No. 5
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.5.2156–2166.2001

Comparison of Statistical Methods for Identification of *Streptococcus thermophilus, Enterococcus faecalis,* and *Enterococcus faecium* from Randomly Amplified Polymorphic DNA Patterns

GIANCARLO MOSCHETTI,[1] GIUSEPPE BLAIOTTA,[1] FRANCESCO VILLANI,[1] SALVATORE COPPOLA,[1] AND EUGENIO PARENTE[2]*

Dipartimento di Scienza degli Alimenti, Università degli Studi di Napoli "Federico II," 80055 Portici,[1] and Dipartimento di Biologia, Difesa, e Biotecnologie Agro-Forestali, Università degli Studi della Basilicata, 85100 Potenza,[2] Italy

A supervised artificial neural network for the discrimination of RAPD-PCR patterns

TABLE 3. Cross-tabulation matrix (true identification in rows, predicted identification in columns) for identification of the strains listed in Table 1 with LDA or BN

Method	Organism	No. of strains of:				Total no. of strains	% Correct identifications
		EFM	EFS	ST	OTH		
LDA[a]	EFM	7	3	0	1	11	64
	EFS	0	24	0	1	25	96
	ST	1	6	72	0	79	91
	OTH	10	7	0	6	23	26
BN[b]	EFM	9	0	0	2	11	82
	EFS	0	24	0	1	25	96
	ST	0	0	75	4	79	95
	OTH	3	0	1	19	23	83

[a] A strain was scored as belonging to the other species (OTH) group if the probability for identification as *E. faecium* (EFM), *E. faecalis* (EFS), and *S. thermophilus* (ST) was <0.80.

[b] A strain was scored as belonging to the other species group if the output for the winning node (i.e., the node with the lowest output) was >0.20.

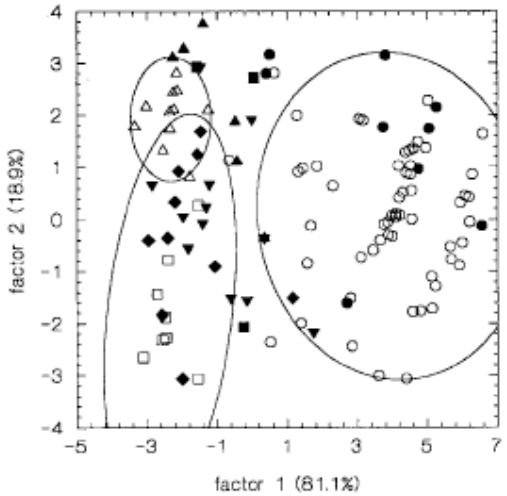


FIG. 4. Canonical score plot of simplified RAPD-PCR patterns obtained with primer XD9 for 138 strains of thermophilic streptococci. The canonical scores were calculated by discriminant analysis for the identification of *S. thermophilus* (○), *E. faecalis* (△), and *E. faecium* (□) using RAPD-PCR patterns for a set of 93 strains (Table 1, group a). Other symbols: ♦, *Streptococcus* spp.; ▼, other enterococci. Open symbols correspond to patterns used for building the model; closed symbols correspond to patterns not used for building the model. The 95% confidence ellipses for the patterns of each species used for building the model are also shown.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2001, p. 2156–2166 Vol. 67, No. 5
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.5.2156–2166.2001

Comparison of Statistical Methods for Identification of *Streptococcus thermophilus, Enterococcus faecalis,* and *Enterococcus faecium* from Randomly Amplified Polymorphic DNA Patterns

GIANCARLO MOSCHETTI,[1] GIUSEPPE BLAIOTTA,[1] FRANCESCO VILLANI,[1] SALVATORE COPPOLA,[1] AND EUGENIO PARENTE[2]*

Dipartimento di Scienza degli Alimenti, Università degli Studi di Napoli "Federico II," 80055 Portici,[1] and Dipartimento di Biologia, Difesa, e Biotecnologie Agro-Forestali, Università degli Studi della Basilicata, 85100 Potenza,[2] Italy

A supervised artificial neural network for the discrimination of RAPD-PCR patterns

TABLE 2. Performance of a supervised ANN (BN), LDA, and CT for the identification of *S. thermophilus*, *E. faecalis*, and *E. faecium* using simplified RAPD-PCR patterns obtained with primer XD9

No. (%) of patterns in the training set:	Median (range) % correct identifications obtained with[a]:		
	BN	LDA	CT
169 (90)	100 (100–100)	100 (100–100)	96 (94–100)
158 (80)	100 (100–100)	100 (95–100)	97 (94–98)
132 (70)	100 (100–100)	95 (94–100)	98 (95–98)
113 (60)	100 (100–100)	96 (96–98)	96 (87–97)
94 (50)	99 (97–100)	95 (94–99)	93 (88–98)
75 (40)	97 (96–100)	91 (84–96)	96 (94–97)

[a] Values are for five replicate runs.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 2001, p. 2156–2166 Vol. 67, No. 5
0099-2240/01/$04.00+0 DOI: 10.1128/AEM.67.5.2156–2166.2001

Comparison of Statistical Methods for Identification of *Streptococcus thermophilus*, *Enterococcus faecalis*, and *Enterococcus faecium* from Randomly Amplified Polymorphic DNA Patterns

GIANCARLO MOSCHETTI,[1] GIUSEPPE BLAIOTTA,[1] FRANCESCO VILLANI,[1] SALVATORE COPPOLA,[1] AND EUGENIO PARENTE[2]*

Dipartimento di Scienza degli Alimenti, Università degli Studi di Napoli "Federico II," 80055 Portici,[1] and Dipartimento di Biologia, Difesa, e Biotecnologie Agro-Forestali, Università degli Studi della Basilicata, 85100 Potenza,[2] Italy

Some rights reserved