Vol. 195: 47–59, 2000 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 31

# Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data

**Lynne Boddy[1,*], C. W. Morris[2], M. F. Wilkins[1], Luan Al-Haddad[2], G. A. Tarran[3], R. R. Jonker[4], P. H. Burkill[3]**

[1]Cardiff School of Biosciences, University of Cardiff, Cardiff CF10 3TL, United Kingdom
[2]School of Computing, University of Glamorgan, Pontypridd, United Kingdom
[3]Centre for Coastal and Marine Sciences, Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, United Kingdom
[4]AquaSense Lab, Kruislaan 411, 1090 HC Amsterdam, The Netherlands

ABSTRACT: Radial basis function artificial neural networks (ANNs) were trained to discriminate between phytoplankton species based on 7 flow cytometric parameters measured on axenic cultures. Comparison was made between the performance of networks restricted to using radially-symmetric basis functions and networks using more general arbitrarily oriented ellipsoidal basis functions, with the latter proving significantly superior in performance. ANNs trained on 62, 54 and 72 taxa identified them with respectively 77, 73 and 70% overall success. As well as high success in identification, high confidence of correct identification was also achieved. Misidentifications resulted from overlap of character distributions. Improved overall identification success can be achieved by grouping together species with similar character distributions. This can be done within genera or based on groupings indicated in dendrograms constructed for the data on all species. When an ANN trained on 1 data set was tested with data on cells grown under different light conditions, overall successful identification was low (<20%), but when an ANN was trained on a combined data set identification success was high (>70%). Clearly it is essential to include data on cells covering the whole spectrum of biological variation. Ways of obtaining data for training ANNs to identify phytoplankton from field samples are discussed.

KEY WORDS: Radial basis functions · Neural networks · Principal component analysis · Dinoflagellates · Prymnesiomonads · Flagellates · Cryptomonads · Diatoms

## INTRODUCTION

Phytoplankton play a pivotal role in marine ecosystems — collectively fuelling the food web, sometimes forming nuisance blooms, and being implicated in climate control. Knowledge of their population dynamics, distribution and abundance in the world's oceans is crucial. There is therefore a need for a technique capable of providing detailed descriptions of the species composition of phytoplankton populations from water samples. Research has been hampered by the limitations of traditional identification and enumeration techniques. Microscopic analysis in the laboratory is laborious and time-consuming, abundance estimates are uncertain due to limitations on the number of cells that can be counted, and interesting phenomena cannot be followed up directly because analysis is often performed a long time after sampling. The use of image analysis is one possibility, and has been used successfully to discriminate 23 dinoflagellate species (Culverhouse et al. 1996). It is, however, computationally intensive. HPLC has been used as a chemotaxonomic technique for *bulk* samples, but it has limited use as a diagnostic tool because it cannot provide fine resolution of taxa (Jeffrey et al. 1997). It is also slow.

Analytical flow cytometry (AFC) may provide a solution to this problem. Light scatter, diffraction and fluorescence parameters are measured on individual cells, at rates of up to $10^3$ cells $s^{-1}$ (Burkill & Mantoura 1990), pro-

*E-mail: boddyl@cardiff.ac.uk

ducing sets of characteristic 'signature' data patterns (1 for each cell) which may allow taxa to be discriminated. The use of a sorter module allows individual cells, for which the data pattern satisfies selected criteria, to be collected for further culture and microscopic analysis — a significant advantage over the other techniques.

AFC has already proved a valuable research tool (Jonker et al. 1995), but its potential cannot be fully realised until appropriate ways of analysing the vast quantities of multivariate data that it generates have been developed. Commonly, bivariate scatter plots of one flow cytometric parameter against another are still utilised (e.g. Hofstraat et al. 1991, Jonker et al. 1995), but this loses much of the information content of the signatures. Multivariate statistical methods have been applied (e.g. Demers et al. 1992, Carr et al. 1996); while these can work well where the data distribution can be approximated by a simple parametric model, this is often not the case for AFC data, which are frequently multimodal. Non-parametric statistical density estimation methods such as Parzen windows and *k*-nearest neighbours (Schalkoff 1992) can overcome this but are computationally intensive, posing problems if the result of the analysis is to be used to drive a real-time cell sorter module.

An extremely powerful alternative is to employ artificial neural networks (ANNs) (Fu 1994, Haykin 1994), which are both non-parametric and computationally efficient in use. ANNs were first developed to mimic the storage and analytical operations of the brain. (Detailed treatment is provided by, for example, Caudill & Butler 1990, Boddy & Morris 1999 [both non-mathematical], Schalkoff 1992, Hush & Horne 1993, Haykin 1994 and Fu 1994.) They are not rule-based, but rather they 'learn' or 'train' from examples presented to them. Essentially, there are 2 types of training — supervised and unsupervised. With the latter, patterns are presented to the network and it forms its own groupings of the data. In contrast, with supervised training, which is appropriate for identification, data patterns (in this case flow cytometric signatures) of *known* identity are presented to the ANN as exemplars. Once trained, any data pattern can be presented to the ANN and the output analysed to find the most likely identity of that pattern. ANNs have been successfully used to analyse flow cytometric data (e.g. Frankel et al. 1989, 1996, Balfoort et al. 1992, Morris et al. 1992, Smits et al. 1992, Boddy et al. 1994, Wilkins et al. 1994a,b, 1996), but, apart from 1 study (Boddy et al. 1994), only a few taxonomic categories have been discriminated. Scaling up is not a trivial task. We examine the issues involved in the application of a particular ANN type, the radial basis function (RBF) network, for the discrimination of up to 72 phytoplankton species. RBF ANNs have been at least as successful as other types in analysis of biological data (Wilkins et al. 1994b, 1996, Morgan et al. 1998). Moreover, they train rapidly and detect 'novel' patterns, for which the identity is not known to the network (Morris & Boddy 1996).

## RADIAL BASIS FUNCTION ARTIFICIAL NEURAL NETWORKS

RBF ANNs are composed of 3 interconnected layers of 'nodes', analogous to neurons (Fig. 1): an 'input layer' containing 1 node per character (in this case AFC parameter), a 'hidden layer', and an 'output layer' containing 1 node per possible identity (in this case corresponding to biological taxa).

A data pattern is presented to the input layer, which serves merely to distribute input data to the hidden layer. Each hidden layer node (HLN) represents a separate basis function (a function for which the value depends solely on the distance between the input data pattern and a fixed point, termed the basis function centre). The basis function centres are collectively positioned so as to represent the distribution of the data patterns throughout the data space. The distances between the input data pattern and the basis function centres are defined by a distance metric, which determines the shape of the basis functions. The Euclidean

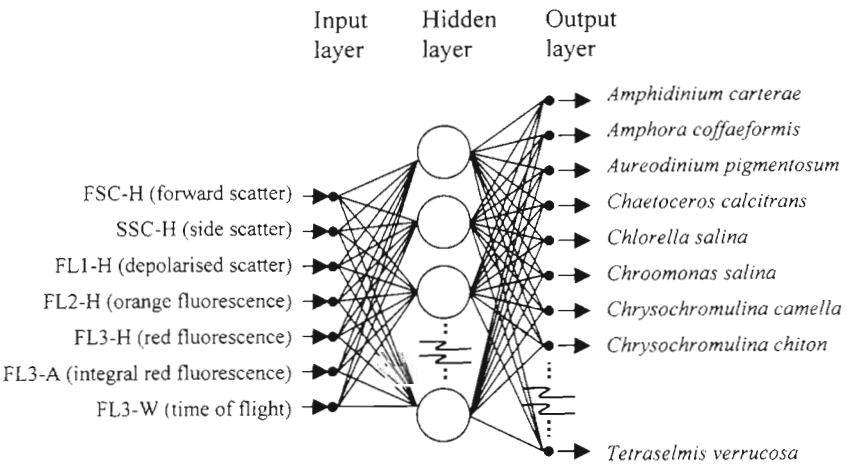


Fig. 1. Schematic of a radial basis function artificial neural network comprising an input layer with 7 nodes (1 per flow cytometric parameter), a hidden layer and an output layer with 1 node per taxon to be identified

distance metric produces hyperspherical (radially symmetric) basis functions about the basis function centres. By independently scaling each dimension of the data, these generalise to hyperellipsoidal (non-radially-symmetric) basis functions for which the principal axes are constrained to lie along the axes of the data space. The Euclidean distance metric is a restricted form of the more general but significantly more computationally intensive Mahalanobis distance metric, which allows the hyperellipsoids to adopt any orientation that best fits the data distribution (Haykin 1994). The initial locations of the basis function centres may be randomly chosen, or be the result of some form of clustering algorithm, e.g. learning vector quantisation (Kohonen 1990). The spatial extent of each basis function may either be constant or be determined by the data. The number of HLNs can be found automatically by starting with a large number of candidate HLNs and selecting from this an optimal subset, e.g. using an orthogonal least squares algorithm (Chen et al. 1991).

The response of all of the hidden layer basis functions is combined by the output layer to form *a posteriori* estimates of the likelihood that the given input pattern belongs to each of the taxa known to the network (Richard & Lippmann 1991). Each output layer node corresponds to a different possible identity, and the most likely identity is found by selecting the output layer node with the highest output value. The decision boundaries formed between taxa (along which the 2 most likely taxa are equally probable) can be arbitrarily complex, depending on basis function locations, number and size.

## METHODS

**Phytoplankton cultures.** Data were collected on 2 separate occasions during the course of 2 marine flow cytometry projects, giving rise to 2 independent data sets denoted A (containing 61 species, 1 of which [*Emiliania huxleyi*] was present as 2 strains) and B (containing 54 species) (see Tables 1 & 2). Forty-three species were common to both data sets. Together the species cover a wide range of morphologies and sizes (approx. 1 to 45 µm) representative of natural nanophytoplankton/flagellate communities in Northern European seas. Phytoplankton cultures, obtained from the Plymouth Culture Collection (Marine Biological Association, UK) and the Alfred Wegener Institute (Bremen, Germany), were maintained at 15°C (± 1°C) and were illuminated on a 12:12 h light:dark cycle at 50 (A) or 130 (B) µmol quanta $m^{-2}$ $s^{-1}$. Batch cultures were grown for several weeks before analysis in 250 ml conical flasks (A) or in 1 l polycarbonate bottles (Nalgene™) (B), and were sub-cultured every 3 to 4 d to maintain cultures in exponential growth. F/10 medium was generally used for culturing, although some cultures were grown in F/2 medium (Guillard & Ryther 1962), with or without soil extract.

**Flow cytometric analysis.** All cultures were analysed by flow cytometry (AFC) using a Becton Dickinson FACSort™ flow cytometer equipped with a vertically polarized 15 mW argon ion laser emitting blue light at 488 nm and FACStation™ acquisition and analysis software, and using instrument settings that had previously been found to allow good discrimination of the type of particle encountered in plankton analysis. Data acquisition was triggered on chlorophyll fluorescence using laboratory cultures of *Micromonas pusilla* (1 to 3 µm) to set the lower analysis threshold. The flow cytometer detector array consisted of 2 fluorescence photo-multiplier tubes (PMTs), 2 light scatter PMTs and a photodiode for forward light scatter. For each particle detected by the cytometer, measurements were made for cellular forward light scatter, integrated and peak chlorophyll fluorescence (>650 nm), the width of the chlorophyll fluorescence pulse or time-of-flight (a measure of particle length), peak phycoerythrin fluorescence (585 ± 21 nm), and side scatter and depolarised light scatter (to enhance the discrimination of coccolithophores). Each measurement of these 7 parameters collectively forms a data pattern characterising an individual cell (or chain/aggregate of cells). Samples were run for 4 min at a flow rate of 100 ± 6 µl $min^{-1}$, with analogue signals from the detectors being digitally converted and stored on computer as listmode data. Instrument drift was monitored by analysing Coulter Flowset calibration particles several times each day.

**Software.** The software used comprised 2 applications developed during the AIMS (Automated Identification and Characterisation of Microbial Populations) project, running on a Pentium PC under Windows95. CytoWave is a flow cytometric data visualisation program, allowing the data to be displayed on multiple 2-D dotplots, and clusters of events in the data to be defined and excluded if desired. AimsNet is a multivariate data analysis program incorporating RBF ANNs, allowing ANNs to be trained to discriminate between selected phytoplankton species and groups of species. The 2 applications are closely integrated, allowing the user to select data within CytoWave, pass it to a trained ANN within AimsNet for neural network analysis, and display the results of the analysis superimposed on the original data.

**Preprocessing cytometric data.** The flow cytometry data for each selected culture was 'gated' using CytoWave to remove any clusters of events originating from 'noise particles' such as inorganic particles, bacterial

contaminants, cellular debris, etc. This was generally achieved by omitting all events with low red fluorescence signals, since such 'noise' particles contain no photosynthetic pigments. The data for some cultures were multimodal, reflecting the presence of clumps of 2 or more cells and cells at different stages of development.

Before presentation to the network the data were linearly rescaled. This procedure is commonly required by neural networks to ensure that equal emphasis is placed by the network on each input parameter when forming an identification; were this not done, the network would tend to make most use of the parameter with the largest absolute range of values, while ignoring the rest, even if they contained useful discriminatory information. In fact, the absolute signal intensities contain no information, since most flow cytometric parameters are measured in arbitrary units that depend on factors such as instrument settings and the optical alignment. The distribution of the training data set was analysed, and a linear transformation calculated such that the distribution of each parameter of the training data set after transformation had a mean of 0.0 and a standard deviation of 1.0. This transformation was subsequently applied to all data presented to the network.

**Training and testing procedures for RBF ANNs.** AimsNet was used to train RBF ANNs to discriminate between the species in both data sets and in a combined data set (see below). All networks were trained using 500 randomly selected data patterns for each species (i.e. data from 500 randomly selected individuals of that species).

The training procedure started by defining 6 candidate HLNs to represent each of the species. The basis function centres were positioned using Kohonen learning-vector quantisation (Kohonen 1990). The spatial extents of the basis functions were determined by allocating each data pattern of the training data to the closest basis function centre, and calculating the covariance matrix of the cluster of patterns allocated to each basis function; the inverse of this matrix is then used in the calculation of the Mahalanobis distance for that basis function. An optimal subset of these HLNs was selected by means of the orthogonal least-squares elimination technique (Chen et al. 1991). The output layer weights were then calculated using matrix inversion. Finally the network performance was optimised to reduce the network output error by 10 iterations of a conjugate directions gradient descent learning procedure; similar procedures have been shown to significantly improve recognition performance (Wettschereck & Dietterich 1992).

Once trained, the networks were tested using an independent set of 500 randomly selected data patterns for each species, and the results recorded in a 'misidentification matrix' **M**, the elements $m_{ij}$ of which indicated the proportion of test patterns for taxon $i$ that were identified by the network as belonging to taxon $j$. From this misidentification matrix, 2 sets of probabilities were recorded for each taxon: (1) the *probability of correct identification* of a taxon; (2) the *identification confidence* of a taxon. The former is the *a priori* probability that a randomly selected individual belonging to that taxon will be correctly identified. For taxon $i$ this is estimated by $m_{ii}$, the proportion of correctly identified test patterns from that taxon. The *identification confidence* of a taxon is the likelihood that a randomly selected individual identified as belonging to that taxon really does belong to that taxon, assuming that all taxa are *a priori* equally likely to occur. For taxon $i$ this is estimated by

$$m_{ii} \Big/ \sum_{j=1}^{N} m_{ji}$$

where $N$ is the number of taxa.

**Constructing a dendrogram.** The presence of taxa with overlapping AFC distributions inevitably reduces the identification confidence, because a proportion of identifications made by the network will be wrong. This is not a consequence of any deficiency in the network, but because the AFC data contain insufficient information to completely resolve between taxa. To improve the overall identification confidence (i.e. increase the *a priori* probability that the network's recognition of the pattern will be correct), at the expense of decreased specificity (i.e. the information obtained is less detailed), taxa which cannot be consistently discriminated from one another can be grouped together during training. These patterns would then be identified with less precision than other patterns; however, a reliable indication that a cell belongs to taxon X or Y, but not specifically which of the two, is often preferable to an unreliable identification as, say, taxon X.

The misidentification matrix can be analysed to produce a dendrogram (e.g. Fig. 2) that shows the natural order in which the taxa recognised by the network can be grouped together. This can be used as an objective way of finding the groupings of taxa that should be used to achieve a desired level of reliability.

Given 2 groups of taxa, denoted by $G_1$ and $G_2$, we define the mutual misidentification probability (i.e. the probability that a pattern belonging to a taxon in $G_1$ will be misidentified as belonging to a taxon in $G_2$ or vice-versa) by

$$\sum_{i \in G_1} \sum_{j \in G_2} [p(i)p(j|i) + p(j)p(i|j)] \equiv \frac{1}{N} \sum_{i \in G_1} \sum_{j \in G_2} (m_{ij} + m_{ji})$$

where $N$ is the total number of taxa in the combined group $G_1 + G_2$, $p(i)$ is the *a priori* probability of a pattern from taxon $i$ (if all taxa are assumed to be *a priori*

equally likely, this is equal to $1/N$), and $p(j|i) \equiv m_{ij}$ (the element $i, j$ of the misidentification matrix) is the *a priori* probability that a pattern from taxon $i$ will be misidentified as taxon $j$. The indices $i$ and $j$ are over the taxa in $G_1$ and $G_2$ respectively.

At the left hand end of the dendrogram, all the taxa are in separate groups (1 taxon per group). At each stage, the groups which have the highest mutual misidentification probability are merged; this reduces the probability that the network's recognition is wrong

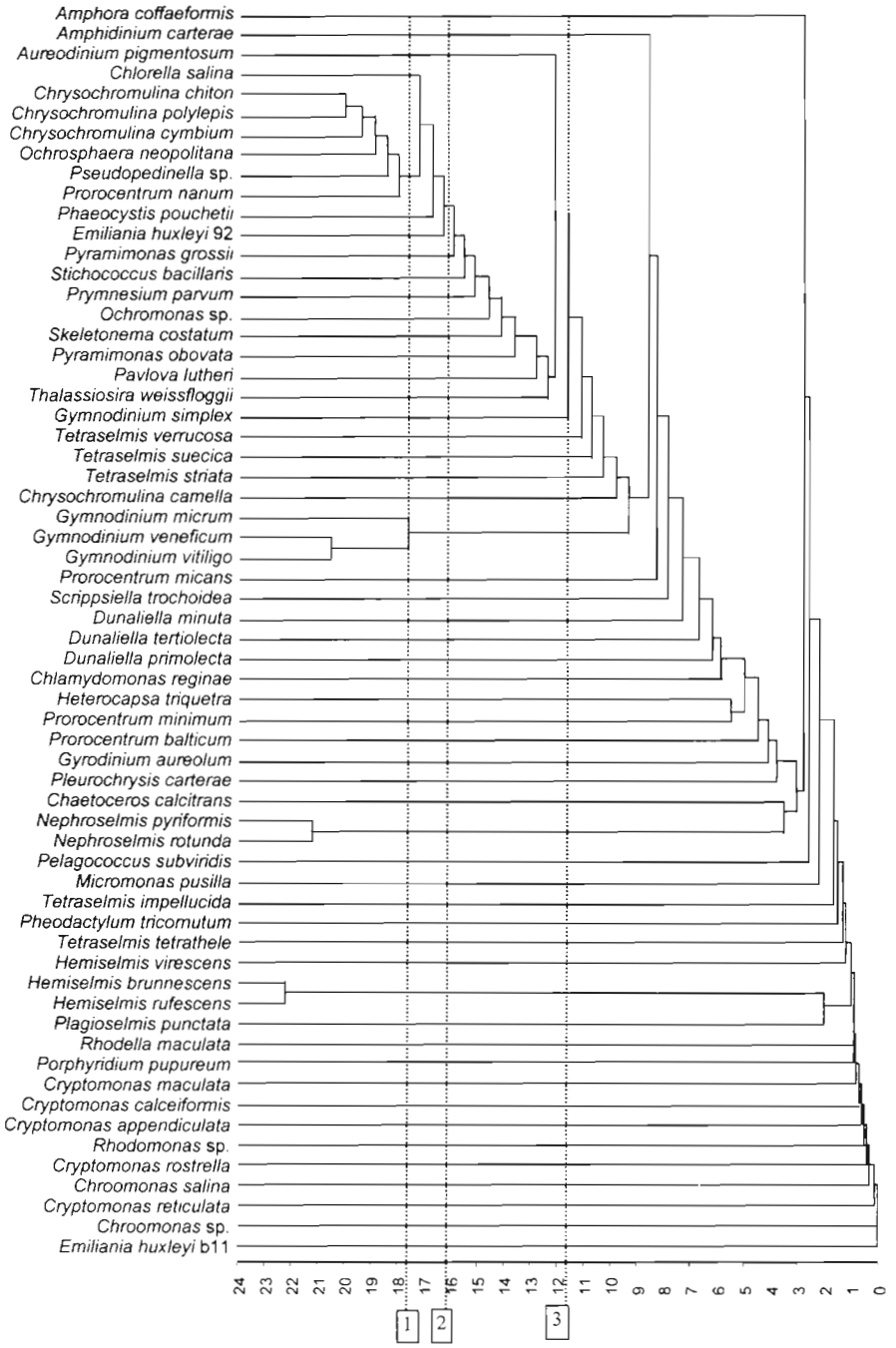


Fig. 2. Dendrogram showing the order in which taxa were clustered by applying the method described in the text to the results matrix of a Mahalanobis network trained on data set A. Clustering proceeds from left to right. Initially each taxon is in a separate group. At each clustering stage, the 2 groups of taxa with the highest mutual confusion are merged, until only 1 group remains. The ordinate axis shows the percentage of misidentified data remaining at each clustering stage, while the dotted lines show the positions corresponding to the 3 groupings referred to in Table 4, with 40, 50 and 54 groups respectively

Table 1. Comparison of the performance of RBF networks trained using Mahalanobis and Euclidean distance metrics, using data set A. For each species, the mean percentage identification and corresponding standard error of the mean (SEM) is shown, measured over 5 replicate trials. The numbers of HLNs ranged from 127 to 154 for the Euclidean networks, and from 135 to 146 for the Mahalanobis networks. Overall performance was 77.3% with SEM 0.16% (Mahalanobis) and 73.1% with SEM 0.31% (Euclidean)

| Group | Class | Species | Size (µm) | Mahalanobis % correctly identified | Mahalanobis SEM | Euclidian % correctly identified | Euclidian SEM |
|---|---|---|---|---|---|---|---|
| Cryptophytes | Cryptophyceae | *Chroomonas* sp. | 8–10 | 95 | 0.2 | 91 | 0.8 |
| | | *Chroomonas salina* | 5–12 | 92 | 0.1 | 86 | 1.0 |
| | | *Cryptomonas appendiculata* | 15–25 | 98 | 0.1 | 97 | 0.7 |
| | | *Cryptomonas calceiformis* | 10–15 | 93 | 0.4 | 92 | 0.3 |
| | | *Cryptomonas maculata* | 12–20 | 91 | 0.4 | 89 | 1.7 |
| | | *Cryptomonas reticulata* | 18–25 | 95 | 0.3 | 94 | 0.4 |
| | | *Cryptomonas rostrella* | 16–25 | 99 | 0.0 | 99 | 0.0 |
| | | *Hemiselmis brunnescens* | 5–8 | 65 | 1.1 | 47 | 2.2 |
| | | *Hemiselmis rufescens* | 4–9 | 59 | 2.0 | 66 | 1.8 |
| | | *Hemiselmis virescens* | 5–8 | 96 | 0.2 | 95 | 0.2 |
| | | *Plagioselmis punctata* | 6–9 | 92 | 0.7 | 84 | 1.3 |
| | | *Rhodomonas* sp. | 8–13 | 92 | 0.8 | 88 | 0.8 |
| Flagellates | Prasinophyceae | *Micromonas pusilla* | 1–3 | 99 | 0.1 | 96 | 0.6 |
| | | *Nephroselmis pyriformis* | 4–7 | 70 | 1.4 | 68 | 1.6 |
| | | *Nephroselmis rotunda* | 6–8 | 51 | 1.8 | 43 | 2.6 |
| | | *Pyramimonas grossii* | 5–10 | 68 | 1.0 | 66 | 1.0 |
| | | *Pyramimonas obovata* | 4–8 | 65 | 0.4 | 64 | 0.4 |
| | | *Tetraselmis impellucida* | 11–19 | 93 | 0.7 | 89 | 2.1 |
| | | *Tetraselmis suecica* | 6–15 | 87 | 0.6 | 81 | 0.7 |
| | | *Tetraselmis verrucosa* | 3–11 | 65 | 2.3 | 42 | 5.8 |
| | | *Tetraselmis tetrathele* | 10–16 | 95 | 0.3 | 95 | 0.1 |
| | | *Tetraselmis striata* | 6–8 | 72 | 1.5 | 73 | 2.2 |
| | Chlorophyceae | *Chlamydomonas reginae* | 11–20 | 91 | 0.4 | 91 | 0.1 |
| | | *Chlorella salina* | 4–8 | 54 | 2.0 | 43 | 2.4 |
| | | *Dunaliella minuta* | 3–12 | 67 | 1.2 | 59 | 1.2 |
| | | *Dunaliella primolecta* | 5–12 | 85 | 0.5 | 83 | 0.9 |
| | | *Dunaliella tertiolecta* | 6–12 | 84 | 0.9 | 80 | 1.2 |
| | | *Stichococcus bacillaris* | 5–8 | 66 | 1.1 | 48 | 4.9 |
| | Rhodophyceae | *Porphyridium pupureum* | 4–6 | 95 | 0.0 | 95 | 0.2 |
| | | *Rhodella maculata* | 7–24 | 94 | 0.5 | 90 | 0.7 |
| | Chrysophyceae | *Ochromonas* sp. | 3–12 | 60 | 1.7 | 52 | 2.1 |
| | | *Pelagococcus subviridis* | 2–3 | 88 | 0.7 | 85 | 0.5 |
| | | *Pseudopedinella* sp. | 8–10 | 74 | 1.1 | 71 | 1.0 |
| Prymnesiophytes | Prymnesiophyceae | *Chrysochromulina camella* | 6–12 | 88 | 0.2 | 85 | 0.3 |
| | | *Chrysochromulina chiton* | 5–9 | 87 | 0.4 | 85 | 0.5 |
| | | *Chrysochromulina cymbium* | 6–10 | 93 | 0.2 | 93 | 0.7 |
| | | *Chrysochromulina polylepis* | 6–8 | 74 | 0.8 | 67 | 0.8 |
| | | *Pleurochrysis carterae* | 10–18 | 92 | 0.6 | 90 | 0.5 |
| | | *Emiliania huxleyi* b11 | 5–7 | 86 | 0.4 | 84 | 0.3 |
| | | *Emiliania huxleyi* 92 | 5–6 | 62 | 0.7 | 59 | 0.6 |
| | | *Ochrosphaera neopolitana* | 8–10 | 43 | 1.3 | 33 | 2.1 |
| | | *Pavlova lutheri* | 4–6 | 61 | 1.3 | 59 | 1.6 |
| | | *Phaeocystis pouchetii* | 3–6 | 90 | 1.0 | 83 | 0.9 |
| | | *Prymnesium parvum* | 8–10 | 97 | 0.1 | 96 | 0.2 |
| Diatoms | Bacillariophyceae | *Amphora coffaeformis* | 10–20 | 81 | 0.8 | 80 | 0.9 |
| | | *Chaetoceros calcitrans* | 4–6 | 42 | 1.5 | 43 | 0.8 |
| | | *Phaeodactylum tricornutum* | 8–35 | 78 | 0.5 | 73 | 0.8 |
| | | *Skeletonema costatum* | 3–5 | 61 | 0.7 | 57 | 1.3 |
| | | *Thalassiosira weissflogii* | 12–20 | 80 | 0.6 | 76 | 1.3 |
| Dinoflagellates | Dinophyceae | *Amphidinium carterae* | 15–20 | 75 | 0.9 | 68 | 1.4 |
| | | *Aureodinium pigmentosum* | 7–12 | 87 | 0.6 | 85 | 0.3 |
| | | *Gymnodinium micrum* | 8–15 | 72 | 1.1 | 65 | 4.2 |
| | | *Gymnodinium simplex* | 6–10 | 64 | 1.8 | 63 | 2.0 |
| | | *Gymnodinium veneficum* | 9–16 | 44 | 2.2 | 23 | 3.8 |
| | | *Gymnodinium vitiligo* | 7–22 | 68 | 1.3 | 71 | 0.7 |
| | | *Gyrodinium aureolum* | 35–45 | 86 | 0.7 | 87 | 1.3 |
| | | *Heterocapsa triquetra* | 15–27 | 76 | 1.1 | 74 | 0.5 |
| | | *Prorocentrum balticum* | 9–15 | 71 | 1.1 | 62 | 2.1 |
| | | *Prorocentrum micans* | 30–40 | 80 | 0.3 | 79 | 0.8 |
| | | *Prorocentrum minimum* | 16–18 | 60 | 1.0 | 57 | 1.3 |
| | | *Prorocentrum nanum* | 8–10 | 56 | 1.9 | 53 | 1.9 |
| | | *Scrippsiella trochoidea* | 30–42 | 51 | 2.1 | 45 | 3.0 |

Table 2. Percentage correct identification of each species (% corr.), and the corresponding percentage confidence that identification is correct (% conf.), for Mahalanobis RBF networks trained on data set A (62 taxa), data set B (54 taxa), and the combined data set A+B (72 taxa), with 135, 132 and 147 HLNs respectively. An asterisk indicates that one of the networks identified the species with very different success from the other networks. The final column shows for each species in data set B any species which it was misidentified as on more than 10% of occasions

| Group | Class | Species | A | | B | | A+B | | Misidentified as species from data set B |
|---|---|---|---|---|---|---|---|---|---|
| | | | % corr. | % conf. | % corr. | % conf. | % corr. | % conf. | |
| Cryptophytes | Cryptophyceae | *Chroomonas* sp. | 95 | 98 | | | 93 | 94 | |
| | | *Chroomonas salina* | 93 | 95 | 96 | 94 | 92 | 95 | |
| | | *Cryptomonas appendiculata* | 98 | 95 | 99 | 100 | 96 | 94 | |
| | | *Cryptomonas calceiformis* | 92 | 96 | 96 | 96 | 93 | 92 | |
| | | *Cryptomonas maculata* | 92 | 91 | | | 94 | 90 | |
| | | *Cryptomonas reticulata* | 94 | 98 | 96 | 98 | 95 | 90 | |
| | | *Cryptomonas rostrella* | 99 | 94 | 92 | 79 | 85 | 76 | |
| * | | *Hemiselmis brunnescens* | 65 | 60 | 77 | 77 | 41 | 63 | *P. punctata* (14%) |
| * | | *Hemiselmis rufescens* | 54 | 65 | | | 81 | 68 | |
| | | *Hemiselmis virescens* | 96 | 95 | 95 | 90 | 98 | 83 | |
| | | *Plagioselmis punctata* | 90 | 87 | 79 | 84 | 83 | 75 | *H. brunnescens* (17%) |
| | | *Rhinomonas salina* | | | 95 | 96 | 95 | 98 | |
| | | *Rhodomonas* sp. | 94 | 95 | 91 | 94 | 91 | 92 | |
| Flagellates | Prasinophyceae | *Micromonas pusilla* | 99 | 83 | 79 | 84 | 79 | 73 | *N. atomus* (21%) |
| | | *Nephroselmis pyriformis* | 73 | 60 | 85 | 73 | 64 | 61 | |
| | | *Nephroselmis rotunda* | 46 | 61 | 75 | 68 | 50 | 55 | *Imantonia* sp. (11%) |
| | | *Pyramimonas grossii* | 66 | 71 | 73 | 63 | 58 | 56 | |
| * | | *Pyramimonas obovata* | 65 | 68 | 34 | 54 | 29 | 53 | *E. huxleyi* (10%), *P. grossii* (19%) |
| | | *Tetraselmis impellucida* | 92 | 93 | | | 95 | 94 | |
| * | | *Tetraselmis striata* | 67 | 78 | 28 | 49 | 32 | 71 | *T. suecica* (11%), *T. tetrathele* (14%), *T. verrucosa* (16%) |
| * | | *Tetraselmis suecica* | 86 | 81 | 65 | 68 | 59 | 60 | |
| | | *Tetraselmis tetrathele* | 95 | 91 | 80 | 65 | 94 | 90 | |
| * | | *Tetraselmis verrucosa* | 61 | 76 | 41 | 53 | 84 | 63 | *A. pigmentosum* (15%) |
| | Chlorophyceae | *Chlamydomonas reginae* | 92 | 80 | | | 90 | 75 | |
| * | | *Chlorella salina* | 60 | 58 | 51 | 57 | 25 | 52 | |
| * | | *Dunaliella minuta* | 65 | 75 | 93 | 72 | 45 | 68 | |
| | | *Dunaliella primolecta* | 85 | 83 | 88 | 89 | 87 | 70 | |
| | | *Dunaliella tertiolecta* | 85 | 73 | | | 86 | 81 | |
| | | *Nannochloris atomus* | | | 84 | 77 | 86 | 76 | *M. pusilla* (15%) |
| | | *Stichococcus bacillaris* | 65 | 75 | 39 | 66 | 35 | 51 | |
| | Rhodophyceae | *Porphyridium pupureum* | 95 | 98 | | | 95 | 97 | |
| | | *Rhodella maculata* | 92 | 97 | | | 93 | 98 | |
| | Chrysophyceae | *Ochromonas* sp. | 63 | 56 | 79 | 67 | 51 | 60 | |
| | | *Pelagococcus subviridis* | 89 | 87 | | | 79 | 83 | |
| | | *Pseudopedinella* sp. | 71 | 67 | | | 87 | 67 | |
| Prymnesiophytes | Prymnesiophyceae | *Chrysochromulina camella* | 87 | 72 | 84 | 76 | 77 | 65 | |
| | | *Chrysochromulina chiton* | 62 | 57 | 48 | 58 | 36 | 55 | *C. polylepis* (17%), *P. pouchetii* (11%) |
| * | | *Chrysochromulina cymbium* | 43 | 54 | 88 | 77 | 67 | 60 | |
| | | *Chrysochromulina polylepis* | 59 | 57 | 63 | 56 | 51 | 48 | *C. chiton* (11%), *P. pouchetii* (18%) |
| | | *Chrysotila lamellosa* | | | 53 | 66 | 67 | 59 | *O. neopolitana* (13%) |
| | | *Dicrateria inornata* | | | 46 | 61 | 55 | 57 | *C. rostrella* (12%) |
| | | *Emiliania huxleyi* | 89 | 84 | 83 | 72 | 70 | 81 | |
| | | *Imantonia* sp. | | | 88 | 66 | 85 | 61 | |
| * | | *Ochrosphaera neopolitana* | 38 | 54 | 80 | 63 | 45 | 55 | |
| | | *Pavlova lutheri* | 78 | 71 | 94 | 93 | 84 | 65 | |
| | | *Phaeocystis pouchetii* | 61 | 64 | 61 | 62 | 63 | 50 | *C. polylepis* (18%) |
| * | | *Platychrysis* sp. | | | 43 | 56 | 72 | 51 | *P. parvum* (28%) |
| | | *Pleurochrysis carterae* | 87 | 90 | 97 | 94 | 88 | 86 | |
| * | | *Prymnesium parvum* | 80 | 68 | 61 | 49 | 37 | 52 | *Platychrysis* sp. (13%) |
| Diatoms | Bacillariophyceae | *Amphora coffaeformis* | 88 | 91 | 79 | 69 | 78 | 63 | |
| | | *Chaetoceros affinis* | | | 63 | 64 | 46 | 73 | |
| | | *Chaetoceros calcitrans* | 88 | 84 | 79 | 81 | 80 | 72 | |
| | | *Chaetoceros debilis* | | | 50 | 60 | 45 | 57 | |
| | | *Chaetoceros radicans* | | | 45 | 62 | 48 | 50 | *C. affinis* (10%), *Ochromonas* sp. (10%) |
| | | *Phaeodactylum tricornutum* | 92 | 94 | | | 91 | 72 | |
| | | *Skeletonema costatum* | 72 | 77 | | | 82 | 71 | |
| | | *Surirella* sp. | | | 81 | 78 | 81 | 67 | |
| | | *Thalassiosira weissflogii* | 90 | 73 | | | 92 | 75 | |
| Dinoflagellates | Dinophyceae | *Amphidinium carterae* | 74 | 70 | 91 | 72 | 77 | 66 | |
| * | | *Aureodinium pigmentosum* | 86 | 71 | 44 | 51 | 53 | 67 | |
| * | | *Gymnodinium micrum* | 72 | 62 | 54 | 62 | 54 | 51 | *G. vitiligo* (32%) |
| | | *Gymnodinium simplex* | 63 | 68 | 79 | 66 | 70 | 58 | |
| * | | *Gymnodinium veneficum* | 45 | 61 | | | 81 | 52 | |
| * | | *Gymnodinium vitiligo* | 69 | 61 | 56 | 57 | 11 | 56 | *G. micrum* (24%) |
| | | *Gyrodinium aureolum* | 86 | 88 | | | 89 | 86 | |
| | | *Heterocapsa triquetra* | 78 | 73 | | | 82 | 82 | |
| | | *Prorocentrum balticum* | 75 | 75 | | | 69 | 74 | |
| | | *Prorocentrum micans* | 80 | 59 | 88 | 79 | 73 | 57 | |
| * | | *Prorocentrum minimum* | 57 | 80 | 85 | 81 | 53 | 69 | |
| * | | *Prorocentrum nanum* | 51 | 65 | 51 | 64 | 24 | 70 | |
| | | *Prorocentrum triestinum* | | | 89 | 94 | 93 | 91 | |
| | | *Scrippsiella trochoidea* | 49 | 70 | | | 62 | 67 | |
| Overall % correctly identified | | | 77 | | 73 | | 70 | | |

by the largest amount possible. As the groups are progressively merged the probability that the network's identification is wrong falls towards zero (at the right hand end of the dendrogram, at which point all taxa have been merged into 1 group).

**Comparison of distance metrics.** Ten networks were trained and tested on data set A, 5 using the Euclidean distance metric (allowing for independent scaling of each dimension) and 5 using the Mahalanobis distance metric. All initially had 6 candidate HLNs per taxon, i.e. 372 HLNs total.

**Performance with large numbers of taxa.** Three networks were trained and tested; one on data set A (62 taxa), one on data set B (54 taxa) and one on the combined data set A+B (72 taxa). The latter data set was generated by combining sets A and B together, with data drawn equally from A and B to represent species common to both data sets. All 3 networks used the Mahalanobis distance metric with 4, 4 and 3 candidate HLNs per taxon (totals of 248, 216 and 216 HLNs) respectively.

**Effect of biological variation on identification accuracy.** Two new data sets were created using only the data for the 43 species common to both data sets A and B. These new data sets were denoted A' and B' respectively. Three networks were trained: one on data set A', one on data set B', and one on the combined data set A'+B' generated as above (all containing 43 taxa). All 3 used the Mahalanobis distance metric, initially with 6 candidate HLNs per taxon, i.e. 258 HLNs total. All 3 networks were tested using both A' and B'.

**Effect of grouping taxa on overall identification success.** The test results of one of the networks trained on data set A were analysed and the corresponding dendrogram plotted (Fig. 2). Five further networks were trained and tested on data set A, using the following grouping schemes: grouping species within a genus together if their mutual misidentification probability was greater than 5%, giving 50 taxa; grouping all species within a genus together, giving 37 taxa (genera); and grouping the species in the manner indicated by points 1, 2, and 3 on the dendrogram in Fig. 2, giving 54, 50 and 40 taxa and predicted successful identification rates of 82, 84 and 88% respectively. (For example, at point 1: *Chrysochromulina chiton*, *C. polylepis*, *C. cymbium*, *Ochrosphaera neopolitana*, *Pseudopedinella* sp. and *Prorocentrum nanum* were grouped, while *Gymnodinium veneficum* was grouped with *G. vitiligo*, *Nephroselmis pyriformis* with *N. rotunda*, and *Hemiselmis brunnescens* with *H. rufescens*). All networks used the Mahalanobis distance metric, initially with 6 candidate HLNs per taxon.

Table 3. Percentage correct identification of 3 networks trained on data set A' alone (131 HLNs), data set B' alone (139 HLNs) and the combined data set A'+B' (146 HLNs) (43 species in each case). Each network was tested for its ability to identify the species from both A' and B'

| Species | Network: | 1 | | 2 | | 3 | |
|---|---|---|---|---|---|---|---|
| | Trained: | A' | | B' | | A'+B' | |
| | Tested: | A' | B' | A' | B' | A' | B' |
| *Amphidinium carterae* | | 80 | 4 | 0 | 89 | 73 | 87 |
| *Amphora coffaeformis* | | 91 | 77 | 39 | 85 | 92 | 81 |
| *Aureodinium pigmentosum* | | 87 | 0 | 1 | 46 | 87 | 24 |
| *Chaetoceros calcitrans* | | 90 | 3 | 0 | 80 | 89 | 85 |
| *Chlorella salina* | | 57 | 18 | 0 | 53 | 47 | 52 |
| *Chroomonas salina* | | 96 | 14 | 55 | 97 | 94 | 95 |
| *Chrysochromulina camella* | | 84 | 0 | 9 | 89 | 87 | 70 |
| *Chrysochromulina chiton* | | 63 | 0 | 36 | 50 | 67 | 14 |
| *Chrysochromulina cymbium* | | 41 | 0 | 0 | 91 | 15 | 88 |
| *Chrysochromulina polylepis* | | 63 | 0 | 1 | 64 | 70 | 29 |
| *Cryptomonas appendiculata* | | 99 | 0 | 0 | 99 | 98 | 97 |
| *Cryptomonas calceiformis* | | 94 | 9 | 0 | 96 | 95 | 94 |
| *Cryptomonas reticulata* | | 97 | 0 | 0 | 98 | 97 | 94 |
| *Cryptomonas rostrella* | | 99 | 1 | 1 | 93 | 99 | 87 |
| *Dunaliella minuta* | | 79 | 7 | 9 | 90 | 80 | 83 |
| *Dunaliella primolecta* | | 92 | 0 | 0 | 85 | 90 | 88 |
| *Emiliania huxleyi* | | 79 | 2 | 1 | 86 | 81 | 63 |
| *Gymnodinium micrum* | | 65 | 0 | 0 | 59 | 69 | 60 |
| *Gymnodinium simplex* | | 66 | 1 | 0 | 80 | 61 | 75 |
| *Gymnodinium vitiligo* | | 83 | 0 | 0 | 60 | 82 | 39 |
| *Hemiselmis brunnescens* | | 95 | 32 | 90 | 74 | 90 | 69 |
| *Hemiselmis virescens* | | 96 | 62 | 79 | 93 | 98 | 94 |
| *Micromonas pusilla* | | 99 | 100 | 60 | 100 | 98 | 100 |
| *Nephroselmis pyriformis* | | 72 | 42 | 11 | 85 | 51 | 82 |
| *Nephroselmis rotunda* | | 54 | 8 | 31 | 87 | 47 | 81 |
| *Ochromonas* sp. | | 79 | 61 | 27 | 81 | 75 | 73 |
| *Ochrosphaera neopolitana* | | 57 | 22 | 10 | 78 | 55 | 68 |
| *Pavlova lutheri* | | 78 | 0 | 0 | 94 | 73 | 95 |
| *Phaeocystis pouchetii* | | 65 | 27 | 8 | 55 | 55 | 67 |
| *Plagioselmis punctata* | | 92 | 54 | 20 | 87 | 88 | 81 |
| *Pleurochrysis carterae* | | 96 | 12 | 0 | 98 | 94 | 96 |
| *Prorocentrum micans* | | 83 | 0 | 0 | 85 | 80 | 80 |
| *Prorocentrum minimum* | | 76 | 0 | 90 | 94 | 74 | 89 |
| *Prorocentrum nanum* | | 71 | 2 | 0 | 59 | 60 | 39 |
| *Prymnesium parvum* | | 82 | 10 | 0 | 74 | 70 | 65 |
| *Pyramimonas grossii* | | 71 | 25 | 8 | 70 | 67 | 63 |
| *Pyramimonas obovata* | | 67 | 8 | 4 | 42 | 46 | 37 |
| *Rhodomonas* sp. | | 93 | 6 | 1 | 94 | 96 | 94 |
| *Stichococcus bacillaris* | | 80 | 56 | 20 | 47 | 71 | 47 |
| *Tetraselmis striata* | | 75 | 0 | 0 | 30 | 74 | 14 |
| *Tetraselmis suecica* | | 87 | 5 | 1 | 65 | 90 | 51 |
| *Tetraselmis tetrathele* | | 95 | 1 | 0 | 76 | 94 | 77 |
| *Tetraselmis verrucosa* | | 72 | 2 | 0 | 47 | 70 | 35 |
| Overall % correctly identified | | 80 | 16 | 14 | 77 | 77 | 70 |

## RESULTS

### Comparison of distance metrics

The variation between the 5 replicate optimized networks was low (Table 1). In terms of overall successful identification rate the networks employing the Mahalanobis distance metric consistently outperformed those employing the scaled Euclidean distance metric by about 4%. For individual species there was often little difference, but for *Hemiselmis brunnescens*, *Tetraselmis verrucosa*, *Chlorella salina*, *Stichococcus bacillaris*, *Ochrosphaera neopolitana* and *Gymnodinium veneficum* there was at least 10% and sometimes over 20% greater success with Mahalanobis than with Euclidean distance.

### Performance with large numbers of taxa

Successful identification rate of the network trained on data set A (62 taxa, 135 HLNs) was 77%; that trained on data set B (54 taxa, 132 HLNs) was 73%; and that trained on A+B (72 taxa, 147 HLNs) was 70% (Table 2). Nine, 15 and 23 species respectively were identified with <60% success. When percentage correct identification was high, so too usually was confidence of correct identification. Exceptions include *Gymnodinium veneficum*, with 81% correct identification but only 52% confidence of correct identification, and *Tetraselmis verrucosa*, with 84% successful identification but only 63% confidence of correct identification. Though many species were identified equally well or equally poorly by all of the networks in which they were included, about 18 species were identified with considerably different success by one of the networks (Table 2). Notable examples are: *Gymnodinium vitiligo* identified with 69 and 56% success respectively by the data set A and B networks, but only with 11% success by the network trained on the combined data set A+B; *Aureodinium pigmentosum* with 86% success in the data set A network, but only 44 and 53% success respectively by the networks for data set B and for A+B.

### Effect of biological variation on performance accuracy

When the networks trained on data sets A' and B' (43 taxa) were tested on the data set on which they had been trained, the overall percentage of cells correctly identified was similar, 80 and 77% respectively (Table 3). There were, however, large differences in success for 7 of the species: *Aureodinium pigmentosum*, *Stichococcus bacillaris*, *Tetraselmis striata* and *T. verrucosa* were considerably better identified within data set A', and *Chrysochromulina cymbium*, *Nephroselmis rotunda*, and *Ochrosphaera neopolitana* were better identified in data set B'.

When the same networks were tested on data from the opposite set from that on which they had been trained, overall successful identification for both was less than 20% (Table 3). Nonetheless, a few species were identified well: the network trained on A' could identify *Amphora coffaeformis* (77%) and *Micromonas pusilla* (100%) from B', while that trained on B' could identify *Hemiselmis brunnescens* (90%), *H. virescens* (79%) and *Prorocentrum minimum* (90%) from A'.

When the network trained on the combined data set A' + B' was tested on data from A' and B', overall successful identification was 77 and 70% respectively. However, a quarter of the species were still poorly identified (<70% success); the species poorly identified differed between the 2 sets (Table 3).

### Effect of grouping taxa on overall identification success

Grouping together species that were misidentified as one another improved overall percentage of correct identification (Table 4). Grouping together all species in a genus resulted in an overall success similar to when only species within a genus were grouped if they were considerably misidentified (>5%) with each other, even though the former resulted in 37 groups and the latter in 50 groups. Grouping according to the dendrogram resulted in a success similar to that predicted by the dendrogram (compare Table 4 with Fig. 2).

Table 4. Effect of different groupings of species from data set A on the performance of Mahalanobis RBF networks

| Network | No. of HLNs | No. of groups | Overall % correctly identified |
|---|---|---|---|
| All species separate | 157 | 62 | 76.8 |
| Species within a genus grouped together when percentage misidentified >5% | 147 | 50 | 83.7 |
| All taxa within genus grouped together | 135 | 37 | 84.0 |
| Dendrogram grouping 1 (see Fig. 2) | 154 | 54 | 83.5 |
| Dendrogram grouping 2 (see Fig. 2) | 163 | 50 | 85.6 |
| Dendrogram grouping 3 (see Fig. 2) | 167 | 40 | 88.7 |

## DISCUSSION

This study demonstrates that previous results with a small (12 species) data set (Wilkins et al. 1996) can be scaled to a large number of species. Identification of phytoplankton by RBF networks employing the Mahalanobis distance metric (termed ARBF) are superior to networks using the scaled Euclidean distance, which allows for the different individual variances of the data along each dimension but uses none of the covariance structure. Considerably better identification of some species when using the Mahalanobis distance probably results from the arbitrary orientation of the Mahalanobis basis function better modelling the underlying flow cytometry parameter distributions for these species.

Overall successful identification in excess of 70% for 72 species compares favourably with results of preliminary studies: 84% for 12 species using ARBF (Wilkins et al. 1994b), 92% for 34 species with ARBF (Wilkins et al. 1999), and with 75% for 42 strains (40 species) using a multilayer perceptron (MLP or back propagation) ANN (Boddy et al. 1994). The very high success in the earlier study with 34 species (Wilkins et al. 1999) was attributable to the fact that both marine and freshwater species having different characteristics were used and that 11 flow cytometric parameters were available as opposed to the 7 used here.

The high confidence of correct identification obtained for most species identified successfully is very important. High success in identification of a species is not sufficient alone since other species may also be identified (incorrectly) as that species, giving an overestimate of occurrence in mixed populations.

Species which were successfully identified by networks trained on all 3 data sets clearly have discriminatory flow cytometric 'fingerprints', whereas those which are consistently identified with low success do not. The reasons are less clear cut as to why the successful identification rate of some of the species should differ considerably between the 2 data sets. It is unlikely to be due to any inherent problems with the ANN approach, as variation between replicates was low (Table 1). Occasionally it may simply be due to different positioning of the decision boundaries between species produced by the networks. This is certainly likely to be the case where species not common to both data sets are identified more successfully by the network trained on the combined data set than by the network trained only on the data set in which they occurred, e.g. *Gymnodinium veneficium*, *Hemiselmis rufescens*, *Platychrysis* sp. and *Tetraselmis verrucosa* (Table 2). For example, in the case of *Platychrysis* sp., which is recognised with 43% accuracy by the network trained on B alone but with 72% accuracy by the network trained on A+B, the discrepancy is explained principally by the fact that the network trained on B alone misidentifies 28% of *Platychrysis* sp. as *Prymnesium parvum* (Table 2): the reason for this can be seen by examining the flow cytometric distributions (Fig. 3).

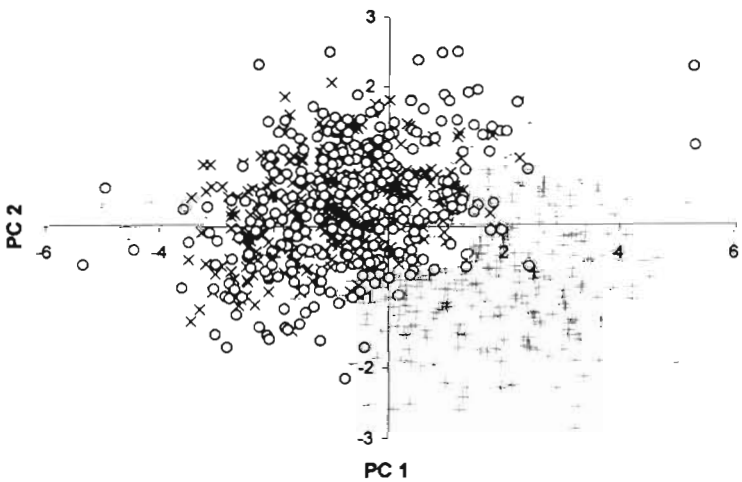


Fig. 3. Distributions of *Prymnesium parvum* from data set A ( ) and data set B (×) and *Platychrysis* sp. from data set B (o), projected on the plane of the first 2 principal components (PC). The distributions of *Platychrysis* sp. and *P. parvum* from data set B overlap considerably. When the distribution for *P. parvum* includes data from data set A (which does not significantly overlap the distribution of *Platychrysis* sp.) the optimal decision boundary is moved and the correct identification rate for *Platychrysis* sp. increases

The network trained on the combined data set moves the decision boundary between the species such that only 3% of *Platychrysis* sp. is misidentified as *P. parvum*; this however means that the proportion of *P. parvum* identified correctly falls from 61 to 37%. Where the identification of a species by the network trained on the combined data set is worse than that by the network trained on either data set alone, the discrepancy may also be explained by species present in the other data set having overlapping character distributions.

For some species common to both data sets, identification rates differed substantially between the data sets (Tables 2 & 3). This is almost certainly due to the same species having different flow cytometric signatures under the different light conditions used for data sets A and B. The low light intensity for A would have resulted in a larger quantity of photosynthetic pigments than for B, yielding higher fluorescence signals. The different flow cytometric signatures can be seen in *Aureodinium pigmento-*

*sum* and *Chrysochromulina camella*, for example, by plotting the first principal component against the second (Fig. 4). With the latter, although the populations have different flow cytometric characteristics, identification success is still high in all networks, presumably because *both* fingerprints are different from those of other species.

When the basis functions have modelled the data well (e.g with an optimized ARBF ANN), misidentifications result from overlap of character distributions. To improve identification success of species whose character distributions overlap, different and/or additional discriminatory characters are required. Grouping together taxa that were misidentified as one another appears to be a good approach to increasing overall successful identification. If it is necessary to discriminate further between species that have been grouped together, and no additional flow cytometric measurements can be obtained, the parameter values that discriminate the group from the rest of the cells in a sample can be used to trigger the flow cytometer's sorting facility. Sorted cells could then be examined using more traditional identification approaches.

The poor performance of networks in making identifications from data sets collected at different times and under different growth conditions highlights a major problem in using this approach for identification of natural mixed populations. Clearly, different populations (i.e. a strain grown under different conditions, or different strains) of a species may have different flow cytometric character distributions (Fig. 4). So long as all of the biological variation is covered in the training data set then good identification can be achieved, as evidenced by the success when training data were selected from both data sets. The high identification success for a few species can be explained by the character distributions (or at least the discriminatory set of characters) remaining similar when the cells were grown under different conditions. For example, *Micromonas pusilla* is easily discriminated because it is considerably smaller than all other species.

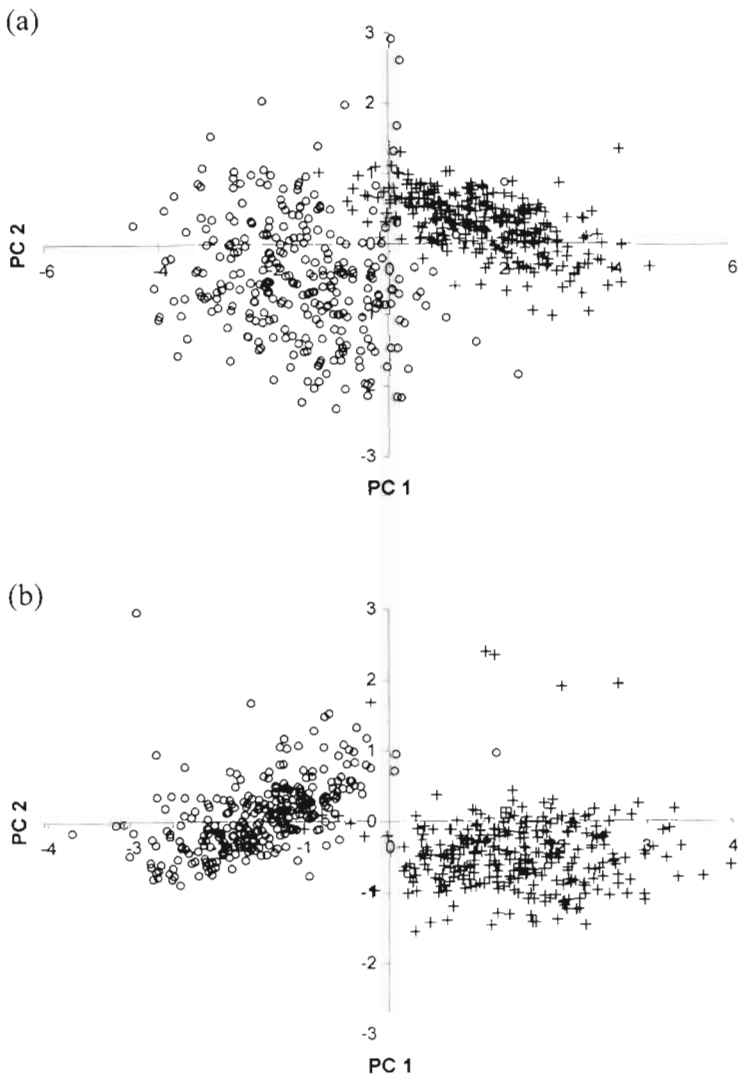


Fig. 4. Plots showing the distributions of (a) *Aureodinium pigmentosum* and (b) *Chrysochromulina camella*, from data set A (+) and data set B (o), projected onto the plane of the first 2 principal components (PC)

## CONCLUSIONS

RBF ANN analysis of flow cytometric data phytoplankton populations provides a powerful quantitative, discriminatory tool. To produce a system capable of identifying field samples, it will be essential to cover the whole spectrum of biological variation within a species encountered in the natural environment in the training data set. It may be possible to achieve this by culturing under a range of conditions, though obtaining training data from actual field samples may be a better alternative. This could be achieved, for example, by performing a statistical (Sneath & Sokal 1973, Dunn & Everitt 1982) or neural (Kohonen 1990, Wilkins et al. 1994a) cluster analysis on a natural sample and then sorting samples into clusters for microscopic identification. Once taxonomic identities can be placed on clusters, these data can form a training set for an ANN, which can subsequently be used for rapid identification of large numbers of cells.

Research into methods for obtaining training data sets from field samples must be a high priority for the future.

A second issue to be addressed is to ensure that the trained ANNs are not tied to any individual cytometer instrument. A solution may be to use standardised calibration beads to define mathematical transformations capable of reducing data captured on specific cytometers, at specific instrument settings, to a standard form. These transformations can then be applied to all data before presentation to the ANN. Any parameters missing from the data will need to be estimated (Boddy et al. 1998).

Finally, it also remains to establish a procedure for converting the results of an ANN analysis of a mixed sample into an estimation of the relative proportions of the different species components, together with reliable confidence limits on the proportion estimates. This work is currently ongoing.

*Acknowledgements.* The AimsNet and CytoWave software were developed during AIMS, a project funded by the Commission of the European Community, CEC grant no. MAS3-CT97-0080. Thanks to all partners for valuable discussion. Flow cytometric measurements were made as part of a Natural Environment Research Council PRiME Special Topic Award (GST/02/1062) (data set A), and during the AIMS project (data set B). We also thank the Alfred Wegener Institute for provision of some cultures.

## LITERATURE CITED

Balfoort HW, Snoek J, Smits JRM, Breedveld LW, Hofstraat JW, Ringelberg J (1992) Automatic identification of algae: neural network analysis of flow cytometric data. J Plankton Res 14:575–589

Boddy L, Morris CW (1999) Artificial neural networks for pattern recognition. In: Fielding AH (ed) Machine learning methods for ecological applications. Kluwer, Dordrecht, p 37–87

Boddy L, Morris CW, Wilkins MF, Tarran GA, Burkill PH (1994) Neural network analysis of flow cytometric data for 40 marine phytoplankton species. Cytometry 15:283–293

Boddy L, Wilkins MF, Morris CW (1998) Effects of missing data on neural network identification of biological taxa: RBF network discrimination of phytoplankton from flow cytometry data. In: Dagli H, Akay M, Buczak CLP, Ersoy AL, Fernandez BR (eds) Intelligent engineering systems through artificial neural networks, Vol 8. American Society of Mechanical Engineers Press, New York, p 655–666

Burkill PH, Mantoura RFC (1990) The rapid analysis of single marine cells by flow cytometry. Philos Trans R Soc A 333: 99–112

Carr MR, Tarran GA, Burkill PH (1996) Discrimination of marine phytoplankton species through the statistical analysis of their flow cytometric signatures. J Plankton Res 18: 1225–1238

Caudill M, Butler C (1990) Naturally intelligent systems. MIT Press, Cambridge, MA

Chen S, Cowan CFN, Grant PM (1991) Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans Neural Networks 2:302–309

Culverhouse PF, Simpson RG, Ellis R, Lindley JA, Williams R, Parisini T, Reguera B, Bravo I, Zoppoli R, Earnshaw G, McCall H, Smith G (1996) Automatic classification of field-collected dinoflagellates by artificial neural network. Mar Ecol Prog Ser 139:281–287

Demers S, Kim J, Legendre P, Legendre L (1992) Analysing multivariate flow cytometric data in aquatic sciences. Cytometry 13:291–298

Dunn G, Everitt BS (1982) An introduction to mathematical taxonomy. Cambridge University Press, Cambridge

Frankel DS, Olson RJ, Frankel SL, Chisholm SW (1989) Use of a neural net computer system for analysis of flow cytometric data of phytoplankton populations. Cytometry 10: 540–550

Frankel DS, Frankel SL, Binder BJ, Vogt RF (1996) Application of neural networks to flow cytometry data analysis and real-time cell classification. Cytometry 23:290–302

Fu LM (1994) Neural networks in computer intelligence. McGraw-Hill, New York

Guillard RRL, Ryther JH (1962) Studies on marine planktonic diatoms. I. *Cyclotella nana* (Hustedt) and *Detonula confervacae* (Cleve) Gran. Can J Microbiol 8:229–239

Haykin S (1994) Neural networks: a comprehensive foundation. Maxwell Macmillan International, New York

Hofstraat JW, de Vreeze MEJ, van Zeijl WJM, Peperzak L, Peeters JCH, Balfoort HW (1991) Flow cytometric discrimination of phytoplankton classes by fluorescence emission and excitation properties. J Fluoresc 1:249–265

Hush DR, Horne BG (1993) Progress in supervised neural networks — what's new since Lippmann? IEEE Sig Proc Mag 10:8–39

Jeffrey SW, Mantoura RFC, Wright SW (eds) (1997) Phytoplankton pigments in oceanography: guidelines to modern methods. UNESCO, Paris

Jonker RR, Meulemans JT, Dubelaar GBJ, Wilkins MF, Ringelberg J (1995) Flow cytometry: a powerful tool in analysis of biomass distributions in phytoplankton. Water Sci Technol 32:177–182

Kohonen T (1990) The self-organising map. Proc IEEE 78: 1464–1480

Morgan A, Boddy L, Morris CW, Mordue JEM (1998) Identification of species in the genus *Pestalotiopsis* from spore morphometric data: a comparison of some neural and non-neural methods. Mycol Res 102:975–984

Morris CW, Boddy L (1996) Classification as unknown by RBF networks: discriminating phytoplankton taxa from flow cytometry data. In: Dagli CH, Akay M, Chen CLP, Fernandez BR, Ghosh J (eds) Intelligent engineering systems through artificial neural networks, Vol 6. American Society of Mechanical Engineers Press, New York, p 629–634

Morris CW, Boddy L, Allman R (1992) Identification of basidiomycete spores by neural network analysis of flow cytometry data. Mycol Res 96:697–701

Richard MD, Lippmann RP (1991) Neural network classifiers estimate Bayesian *a posteriori* probabilities. Neural Comp 3:461–483

Schalkoff RJ (1992) Pattern recognition: statistical, structural and neural approaches. Wiley International, Chichester

Smits JRM, Breedveld LW, Derksen MWJ, Kateman G, Balfoort HW, Snoek J, Hofstraat JW (1992) Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. Anal Chim Acta 258: 11–25

Sneath PHA, Sokal RR (1973) Numerical taxonomy. WH Freeman, San Francisco

Wettschereck D, Dietterich T (1992) Improving the performance of radial basis function networks by learning center locations. Adv Neural Info Process Syst 4:1133–1140

Wilkins MF, Boddy L, Morris CW, Jonker R (1996) A comparison of some neural and non-neural methods for identification of phytoplankton from flow cytometry data. CABIOS 12:9–18

Wilkins MF, Boddy L, Morris CW (1994a) Kohonen maps and learning vector quantization neural networks for analysis of multivariate biological data. Binary 6:64–72

Wilkins MF, Morris CW, Boddy L (1994b) A comparison of radial basis function and backpropagation neural networks for identification of marine phytoplankton from multivariate flow cytometry data. CABIOS 10:285–294

Wilkins MF, Boddy L, Morris CW, Jonker R (1999) Identification of phytoplankton from flow cytometry data using radial basis function neural networks. Appl Environ Microbiol 65:4404–4410

*Editorial responsibility: Otto Kinne (Editor), Oldendorf/Luhe, Germany*

*Submitted: June 22, 1999; Accepted: October 26, 1999*
*Proofs received from author(s): March 17, 2000*