CSIRO PUBLISHING
www.publish.csiro.au/journals/asb *Australian Systematic Botany*, **22**, 31–38

Phyllometric parameters and artificial neural networks for the identification of *Banksia* accessions

Giuseppe Messina[A,B], *Camilla Pandolfi*[A], *Sergio Mugnai*[A,C], *Elisa Azzarello*[A], *Kingsley Dixon*[B] *and Stefano Mancuso*[A]

[A]Department of Horticulture, University of Florence, Viale delle Idee 30, 50019 Sesto Fiorentino (FI), Italy.
[B]School of Plant Biology, University of Western Australia, 35 Stirling Highway, Crawley, WA 6907, Australia.
[C]Corresponding author. Email: sergio.mugnai@unifi.it

Abstract. Taxonomic identification is traditionally carried out with dichotomous keys, or at least computer-based identification keys, often on the basis of subjective visual assessment and frequently unable to detect small differences at subspecies and varietal ranks. The aims of the present work were to (1) clearly discriminate a wide group of accessions (species, subspecies and varieties) belonging to the genus *Banksia* on the basis of 14 phyllometric parameters determined by image analysis of the leaves, and (2) unequivocally identify the accessions with a relatively simple back-propagation neural-network (BPNN) architecture (single hidden layer) in order to develop a complementary method for fast botanical identification. The results indicate that this kind of network could be effectively and successfully used to discriminate among *Banksia* accessions, as the BPNN enabled a 93% unequivocal and correct simultaneous identification. Our BPNN had the advantage of being able to resolve subtle associations between characters, and of making incomplete data (i.e. absence of *Banksia* flower parameters such as the colour or size of styles) useful in species diagnostics. This method is relatively useful; it is easy to execute as no particular competences are necessary, equipment is low cost (scanner connected to a PC and software available as freeware) and data acquisition is fast and effective.

Introduction

Identification of taxa has always been a topic of general interest and often an important task in systematic biological disciplines, such as botany, zoology, microbiology and also palaeontology. Much of this biological identification is still carried out with dichotomous keys, which are classical paper-based kind of expert system that usually must be followed manually (Weeks and Gaston 1997; O'Neill 2007). The main disadvantage is that with a conventional dichotomous key, one needs to hit only one unanswerable couplet, and the identification cannot proceed further. The development of any computer-aided identification key such as computer assisted taxonomy (CAT; Chesmore 2000) is a major advance for taxonomists and non-taxonomists alike because it enables to start the identification process at any point, with characters directly chosen by the user, and to select user's own path through the key. The great advantage of computer-aided identification keys over conventional dichotomous keys is that difficult or missing characters can be ignored. Furthermore, as computer-aided identification keys work through a process of elimination, two questions can sometimes be sufficient for a positive identification.

The partial or complete automation of identification process has been an obvious response when faced with an activity that involves repetitive processes such as taxonomic identification, when the labour costs become too high or when automation offers faster, more replicable or more accurate results. However, the development and application of an automated approach to taxonomic identification has remained relatively unexplored for years (Wheeler 2007). Recently, automation of taxon identification (ATI; Chesmore 2007) on the basis of morphological characters through the capture of digital images of the specimens and data processing via different approaches such as digital automated identification system (DAISY; Gaston and O'Neill 2004) or dedicated algorithms such as artificial neural networks (ANNs) have improved the accuracy of discrimination (Weeks and Gaston 1997; Pandolfi *et al*. 2006; Du *et al*. 2007), mainly owing to their capability of handling incomplete or sterile material.

An ANN is an information-processing paradigm modelled on biological nervous systems, comprising a large number of highly interconnected processing elements (akin to neurons) working in unison to solve specific problems. ANNs are not rule-based, but are trained on examples of the taxa to be identified by an iterative process that alters the internal organisation of the network until it can successfully distinguish between the selected accessions. ANNs have a great potential to partly automate the identification process, especially if coupled with image analysis. This has previously been conducted in biological taxonomy for the identification of bacteria (Giacomini *et al*. 2000; Mouwen *et al*. 2006; Sahin and Aydin 2006), protozoa (Ginoris *et al*. 2007), phytoplankton (Wilkins *et al*. 1999), algae (Balfoort *et al*. 1992; Smits *et al*. 1992), fungi (Morris *et al*. 1992; Morgan *et al*. 1998), insects (Chesmore and Ohya 2004; Van hara *et al*. 2007), spiders (Do *et al*. 1999), molluscs (Hernández-Borges

 11 March 2009 10.1071/SB08003 1030-1887/09/010031

et al. 2004) and fossils (Walsh *et al*. 2007). Moreover, the possibility of using ANNs for plant identification has been recently attempted with a certain success (Clark and Warwick 1998; Mancuso and Nicese 1999; Mancuso *et al*. 1999; Mariño and Tressens 2001; Clark 2004; Mugnai *et al*. 2008).

Banksia (Proteaceae) is an icon Australian taxon, represented by 80 species (George 1981, 1988, 1999). Evidences for a paraphyly of *Banksia* genus with respect to *Dryandra* genus were underlined by Mast and Givnish (2002) and Mast *et al*. (2005). For this reason, Mast and Thiele (2007) proposed new combinations for the species, subspecies and varieties of *Dryandra* to *Banksia. Banksia* species are normally classified on the basis of plant habit, flower and fruit characteristics and leaf shapes. Cladistic analysis based on morphological and anatomical characters have been previously performed on 35 *Banksia* species (Thiele and Ladiges 1996). Traditional taxonomic keys based on leaf parameters and flower characteristics have already been used for the identification of *B. integrifolia* (Thiele and Ladiges 1994); however, there are no known computer-based identification systems for *Banksia*.

This work aims to present a case study of automated identification of 84 plant specimens belonging to the *Banksia* genus, on the basis of leaf morphological features obtained by online image capture and processing. The discrimination of a taxonomically wide selection of species, subspecies and forms, and their unequivocal identification, is carried out by a specific and dedicated ANN in order to test the applicability of this approach as an adjunct to traditional plant-identification methods.

Materials and methods

Plant material

All plant material was collected from Kings Park and Botanic Garden of Perth (WA, Australia, 31°57′41″S, 115°50′22″E) and a living collection located on The Banksia Farm (Mount Barker, WA, Australia, 34°38′15″S, 117°38′58″E). The selected accessions belonged to 67 species and 17 subspecies or varieties of *Banksia* (Table 1), which showed a wide range of leaf sizes and shapes (Fig. 1). The initial sampling planned to collect two distinct sets of leaves for both the training and the validation processes, each one composed of 40 samples per accession. From preliminary tests, 40 was considered the minimum significant number of leaves in a set for each phase of the artificial neural-network (ANN) construction. The training set was used to build the neural model, whereas the validation set was used to verify the correctness of the model obtained with another independent set of samples. Leaves were randomly collected from healthy 1-year-old branches spread across at least four individuals (Table 1) to minimise the effect of the unavoidable leaf-size variability.

Image acquisition and determination of phyllometric parameters

An optical scanner, set at 300 × 300 dpi and 16-million colours, was used to acquire leaf images. In all, 14 phyllometric parameters (nine of them describing the geometrical shape of the leaves and five referring to their different grey levels, Table 2) were automatically determined from each leaf image through an image analysis software (UTHSCSA Image Tool 3.0, http://ddsdx.uthscsa.edu/dig/itdesc.html). In this case, the software needs user guidance only in the first phase of image analysis, to distinguish the leaf outline from the background.

Construction of the neural network

The implemented ANN was a supervised back-propagation neural network (BPNN), an iterative gradient algorithm derived from the multilayer perceptron (MPL). The 14 leaf parameters obtained from the image analysis were used in the neural network as inputs during both the training and the validation phases, whereas the 84 accessions represented the outputs. In our case, the term ‘accession’ included all the collected *Banksia* species, subspecies and varieties, with each accession having its own output node of equal weight. Several hierarchical ANNs were preliminarily created by testing different numbers of hidden layers and nodes per layer. Factors in the hidden layer, such as training scheme, numbers of nodes in the output and input, and connections between them play an important role for the determination of the best configuration (Zurada and Jacek 1992). A suitable ANN architecture with a three-layered structure (input layer, hidden layer and output layer) was searched, by evaluating the dependence of the root mean square (RMS) error on the number of nodes in the hidden layer, as follows:

$$\text{RMS error} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \text{out}_{ij})^2}{N \times M}}, \quad (1)$$

where y_{ij} is the element of the matrix ($N \times M$) for the training set, and out_{ij} is the element of the output matrix ($N \times M$) of the neural network (N is the number of variables in the matrix and M is the number of samples). A single hidden-layer architecture was preferred to a multiple hidden-layers architecture because of its better resistance, robustness and velocity of performance. The overall network architecture (14, n, 84) was chosen, where 14 was the number of nodes (leaf parameters) in the input layer, n was the number of nodes in the hidden layer, and 84 was the number of nodes (accessions) in the output layer. The correct network structure was reached when the RMS error was minimised; after evaluating several nodes in the hidden layer in the range 1–100, the minimal value of RMS error (0.06), with the difference between RMS errors in two consecutive periods less than 1×10^{-4}, was achieved when $n = 50$, so leading to a (14, 50, 84) architecture. This architecture permitted the highest % of correct identification (93% of prediction). The use of a multiple hidden-layers architecture and/or a lower number of hidden nodes did not permit the same accuracy in the identification of the accessions. The training phase was then considered complete, because the ANN achieved the desired statistical accuracy, producing the required outputs for a given sequence of inputs with the best performance.

The output values give an idea about the effectiveness of the ANN in recognising and discriminating the single accessions. In an ideal case, only a class of output, representing the tested accession, would show an average value of 1 (correct identification) while all the other classes would show the value 0 (incorrect identification). This happens occasionally, owing to

Table 1. List of *Banksia* accessions (and their acronyms) collected at Kings Park and Botanic Garden of Perth (WA, Australia) and the Collins' living collection located in Banksia Farm, Mount Barker (WA, Australia)

Numbers indicate the number of plants used for the training (left) and the validation (right) sets for each accession, with 40 leaves picked for each set

B. aemula	AEM	4,4	*B. media*	MED	5,5
B. aculeata	ACU	5,4	*B. media* subsp. *penicillata*	MED P	4,5
B. aquilonia	AQU	5,4	*B. meisneri* subsp. *ascendens*	MEI A	5,5
B. ashbyi	ASH	4,6	*B. meisneri* subsp. *meisneri*	MEI M	5,4
B. attenuata	ATT	4,4	*B. menziesii*	MEN	4,4
B. audax	AUD	6,6	*B. micrantha*	MIC	6,6
B. baueri	BAU	6,5	*B. nutans*	NUT	6,6
B. baxteri	BAX	4,5	*B. oblongifolia* 'Blue flower'	OBL B	5,5
B. benthamiana	BEN	5,5	*B. oblongifolia* 'Spred'	OBL S	5,5
B. brownii 'Intermedia'	BRO I	5,5	*B. occidentalis*	OCC	5,5
B. brownii 'Mountain'	BRO M	5,6	*B. occidentalis* subsp. *formosa*	OCC F	5,5
B. brownii 'Tree'	BRO T	6,6	*B. oligantha*	OLI	4,5
B. burdettii	BUR	5,4	*B. oreophila*	ORE	5,4
B. caleyi	CAL	6,5	*B. ornata*	ORN	4,4
B. candolleana	CAN	5,6	*B. paludosa* subsp. *paludosa*	PAL P	6,6
B. canei	CAE	5,5	*B. paludosa* 'Dwarf'	PAL D	5,5
B. coccinea 'Orange'	COC O	4,4	*B. paludosa* 'Astrolux'	PAL A	5,5
B. coccinea 'Red'	COC R	4,6	*B. pilostylis*	PIL	6,5
B. conferta subsp. *conferta*	CON	5,6	*B. plagiocarpa*	PLA	6,6
B. dryandroides	DRY	6,6	*B. praemorsa* 'Yellow flower'	PRE Y	4,4
B. epica	EPI	6,5	*B. praemorsa* 'Red flower'	PRE R	4,5
B. ericifolia subsp. *macrantha*	ERI	6,4	*B. prionotes*	PRI	6,6
B. ericifolia × *spinulosa*	EXS	4,5	*B. pulchella*	PUL	6,5
B. grandis	GRA	4,4	*B. quercifolia*	QUE	6,6
B. grossa	GRO	5,5	*B. saxicola*	SAX	5,5
B. hookeriana	HOO	5,5	*B. scabrella*	SCA	5,5
B. hookeriana × *prionotes*	HXP	5,5	*B. sceptrum*	SCE	5,6
B. ilicifolia	ILI	6,6	*B. seminuda* 'Red styles'	SEM S	5,4
B. incana	INC	5,6	*B. seminuda* subsp. *remanens*	SEM R	5,4
B. integrifolia shrub form	INT S	5,5	*B. speciosa*	SPE	5,5
B. integrifolia subsp. *compar*	INT C	6,6	*B. seminuda* 'Yellow styles'	SEM Y	6,7
B. integrifolia subsp. *integrifolia*	INT I	6,6	*B. serrata*	SER	6,6
B. integrifolia subsp. *monticola*	INT M	6,7	*B. solandri*	SOL	5,4
B. laevigata subsp. *laevigata*	LEA	4,4	*B. sphaerocarpa* 'Dolichostyla'	SPH	5,5
B. lanata	LAN	4,5	*B. spinulosa* 'Spinulosa'	SPI S	5,5
B. laricina	LAR	4,4	*B. spinulosa* 'Neoanglica'	SPI N	4,5
B. lemanniana	LEM	5,5	*B. spinulosa* 'Cunninghamii'	SPI C	4,4
B. leptophylla 'Leptophylla'	LEP L	6,5	*B. telmatiaea*	TEL	5,5
B. leptophylla 'Melletica'	LEP M	6,6	*B. tricuspis*	TRI	6,5
B. lindleyana	LIN	5,5	*B. verticillata*	VER	5,5
B. littoralis	LIT	4,5	*B. victoriae*	VIC	5,6
B. marginata	MAR	4,4	*B. violacea*	VIO	4,4

the natural variation among leaves, so the output of the expected class often reports a value <1 while the others could show a value >0. In order to measure the likelihood that a species identification was correct, given that the ANN identified an unknown sample as that taxon, the confidence of correct identification (%Conf) was used. %Conf is a parameter identical to the confidence of correct classification used by Morgan *et al*. (1998). It is expressed as a percentage of the proportion of correct identifications with respect to the total number of identifications, as follows (including wrong identifications):

$$\%\text{Conf} = \frac{\text{correct}}{\text{correct} + \text{incorrect}} \times 100. \qquad (2)$$

The trained (14, 50, 84) ANN model was then tested and validated with a validation set of data. A validation test is critical to verify and ensure that the network did not simply memorise the training set but learned the general patterns involved within an application. The validation is a test of prediction power of the model, i.e. its effectiveness in identifying unknown specimens. Moreover, the validation set was used to prevent over-training, i.e. the situation when the model is too complex and training achieves a low error but has a poor generalisation when new samples are processed. In our case, the validation test determined a very high percentage of successful identification by comparing the predicted identity with the known identity of the *Banksia* accessions.

Results

For any accession, the output value of the ANN and the confidence of correct identification (%Conf) are reported in the

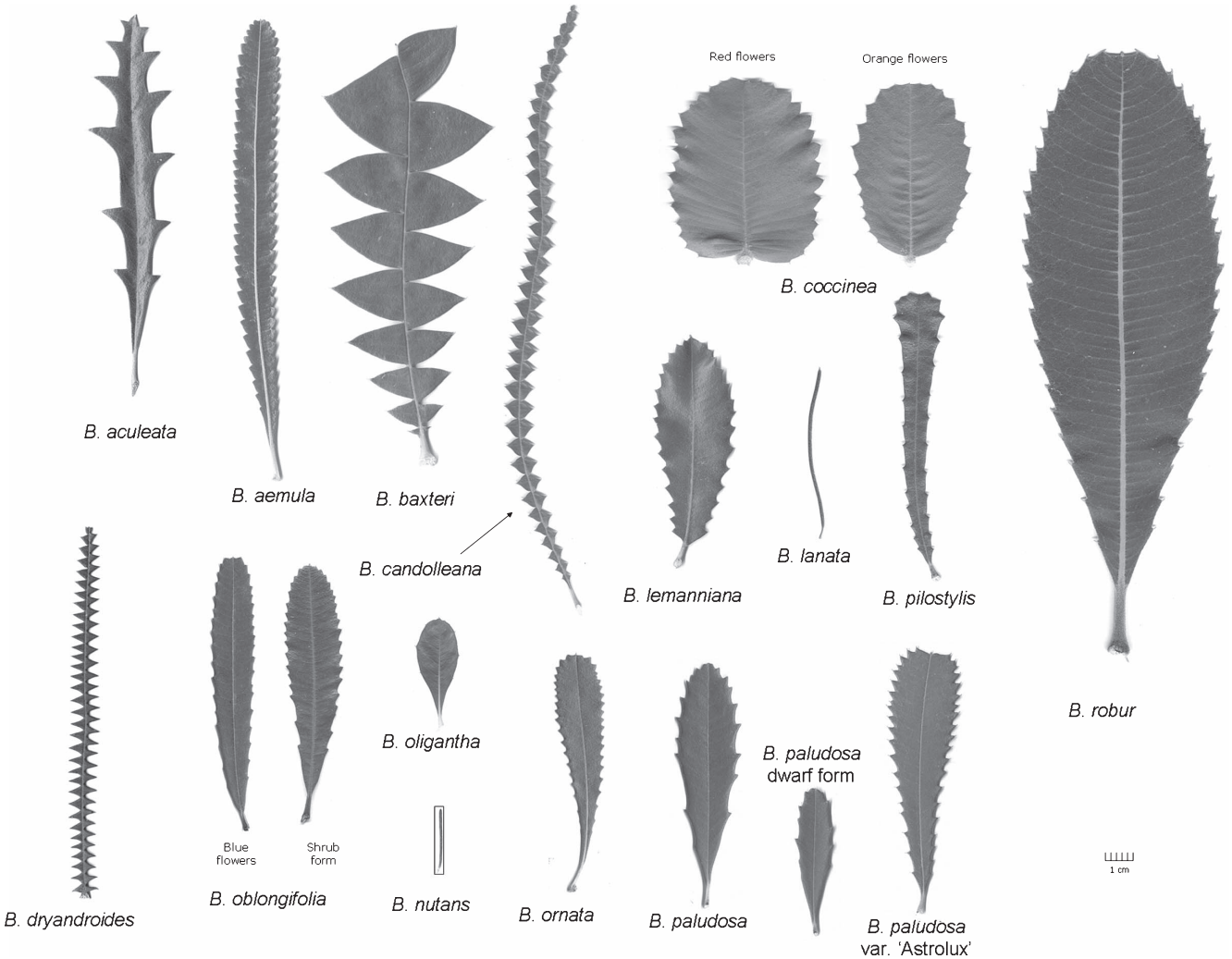

Fig. 1. Leaf shapes belonging to different *Banksia* accessions.

Table 2. Phyllometric parameters used as inputs in the artificial neural networks (ANNs) determined from each leaf by image analysis

No.	Parameter	Definition
1	Area	The area of the leaf
2	Perimeter	The perimeter of the leaf
3	Major axis length	The length of the longest line that can be drawn trough the leaf
4	Minor axis length	The length of the longest line that can be drawn trough the leaf perpendicular to the major axis
5	Roundness	Computed as: $(4 \times \pi \times \text{area})/\text{perimeter}^2$
6	Elongation	The ratio of the length of the major axis to the length of the minor axis
7	Feret diameter	The diameter of a circle having the same area of the leaf
8	Compactness	Computed as sqrt $(4 \times \text{area}/\pi)$/major axis length
9	Integrated density	Computed as the product of the mean grey level and the number of pixels in the image of the leaf
10	Min grey level	Minimum grey level of the leaf
11	Mean grey level	Mean grey level of the leaf
12	Median grey level	Median grey level of the leaf
13	Mode grey level	Mode grey level of the leaf
14	Max grey level	Maximum grey level of the leaf

'misidentification' table (Table 3). The structure of the misidentification table derives from the previously created misidentification matrix (Boddy *et al*. 2000; Clark 2003), in which the matrix rows refer to the species in the test set and the matrix columns are the species to which the test plants are referred by the neural network. In our case, the misidentification

Table 3. Misidentification table, reporting the output values of the artificial neural networks (ANNs), the %Conf and the principal misidentifications for each species

Accession	Output value	%Conf	Principal misidentifications	Accession	Output value	%Conf	Principal misidentifications
ACU	0.76	85.97		MED	0.59	54.10	ILI (0.13); MED P (0.07)
AEM	0.99	94.51		MED P	0.81	86.41	
AQU	0.90	96.95		MEI A	0.94	95.32	
ASH	0.75	86.74		MEI M	1.00	100	
ATT	0.83	84.46		MEN	0.84	87.35	
AUD	0.96	97.02		MIC	0.74	74.89	
BAU	0.71	68.45		NUT	0.64	70.56	TEL (0.13)
BAX	0.96	99.19		OBL B	0.51	47.89	s.e.m. Y (0.23); s.e.m. S (0.13)
BEN	0.84	91.22		OBL S	0.78	86.28	
BRO I	0.57	60.43	BRO M (0.12); BRO T (0.23)	OCC	0.84	85.69	
BRO M	0.84	81.50		OCC F	0.91	88.88	
BRO T	0.74	74.42		OLI	0.79	77.75	
BUR	0.82	77.00		ORE	0.88	93.32	
CAL	0.75	76.20		ORN	0.86	76.34	
CAN	0.96	93.46		PAL A	0.73	69.84	
CAE	0.95	96.97		PAL D	0.74	72.52	
COC O	0.95	96.24		PAL P	0.91	82.04	
COC R	0.96	95.05		PHI	0.52	57.05	ACU (0.12)
CON	0.91	79.77		PLA	0.93	93.69	
DRY	0.94	97.51		PRE R	0.76	72.27	
EPI	0.67	74.23	PRE Y (0.13)	PRE Y	0.58	59.03	EPI (0.14); PRE R (0.13)
ERI	0.98	98.29		PRI	0.96	97.78	
EXS	0.95	93.30		PUL	1.00	100	
GRA	1.00	97.40		QUE	0.93	95.73	
GRO	0.89	81.83		SAX	0.95	100	
HOO	0.89	85.70		SCA	0.85	85.15	
HXP	0.87	85.58		SCE	0.94	87.84	
ILI	0.63	69.28	MED (0.12); PAL A (0.09)	SEM R	0.76	76.95	
INC	0.89	93.53		SEM S	0.62	66.79	s.e.m. Y (0.11); OBL B (0.09)
INT C	0.73	72.94		SEM Y	0.47	43.13	OBL B (0.29); s.e.m. S (0.15)
INT I	0.73	78.86		SER	0.70	78.76	CON (0.07)
INT M	0.65	69.43	INT C (0.11); s.e.m. Y (0.08)	SOL	0.97	100	
INT S	0.91	88.02		SPE	0.87	86.84	
LAN	0.85	78.27		SPH	0.63	66.12	LEP M (0.30)
LAR	0.96	92.22		SPI C	0.94	94.42	
LEA	0.78	72.92		SPI N	0.84	92.21	
LEM	0.75	81.40		SPI S	0.91	87.29	
LEP L	0.58	66.27	LAN (0.11); LEP M (0.06)	TEL	0.51	47.71	VIO (0.25); NUT (0.18)
LEP M	0.61	62.44	SPH (0.28); LEP L (0.07)	TRI	0.95	100	
LIN	0.88	90.26		VER	0.73	73.57	
LIT	0.88	83.97		VIC	0.82	78.01	
MAR	0.90	85.04		VIO	0.60	57.25	TEL (0.24)

matrix was converted into a table because of the high number of accessions. As an example, ANN correctly identified 35 of the 40 unknown *B. menziesii* and therefore had a %Conf value of 87.35%. Table 3 reveals a different effectiveness in the identification process by the ANN. Most of the accessions were well identified through the use of the phyllometric parameters, reaching values higher than 80% for both ANN output and %Conf.

In some cases (*B. aemula*, *B. aquilonia*, *B. audax*, *B. baxteri*, *B. candolleana*, *B. canei*, *B. coccinea* 'Orange', *B. coccinea* 'Red', *B. conferta*, *B. dryandroides*, *B. ericifolia*, *B. grandis*, *B. integrifolia* shrub form, *B. laricina*, *B. meisneri* subsp. *ascendens*, *B. meisneri* subsp. *meisneri*, *B. occidentalis* subsp. *formosa*, *B. paludosa*, *B. plagiocarpa*, *B. prionotes*, *B. pulchella*, *B. quercifolia*, *B. saxicola*, *B. sceptrum*, *B. solandri*, *B. spinulosa*, *B. tricuspis*) the output values were $\geq$0.90, which is considered a threshold value for a complete recognition of the accessions by a ANN (Mugnai *et al*. 2008), so denoting a very high and significant effectiveness in the identification of the selected species. Interestingly, the network was also able to discriminate strongly between some cases of subspecies/varieties/forms (*B. coccinea* 'Red' and 'Orange'; *B. meisneri* subsp. *meisneri* and *ascendens*; *B. brownii* 'Mountain' and 'Tree' forms). Moreover, *B. ericifolia* × *spinulosa* was well identified as a different taxon type (ANN output value of 0.95 and a %Conf of 93.30), with its own peculiar morphological characteristics, and clearly separated from its parental taxa. Conversely, in a few cases the ANN was not able to identify a species completely. This happened when the ANN output value and/or the %Conf of a certain species was approximately 0.5 (50%), and the output

values of other species were concurrently significantly high. Therefore, some cases of misidentification occurred, as shown in Table 3. For example, *B. telmatiaea* (output value 0.51, %Conf 47.71) and *B. oblongifolia* 'Blue flower' (output value 0.51, %Conf 47.89) were not unequivocally distinguished by the ANN, as the output values encompassed other species. For example, ANN confused *B. telmatiaea* with *B. violacea* (output value 0.25) and *B. nutans* (output value 0.25), whereas *B. oblongifolia* 'Blue flower' overlapped the outputs of *B. seminuda* 'Yellow styles' and *B. seminuda* 'Red styles'. The results of the present study indicate that an unequivocal characterisation of *Banksia* species by the use of phyllometric parameters alone is not always possible.

Discussion

The high percentage of correctly identified specimens from the dataset showed that the ANN is practically capable of resolving botanical identification related to *Banksia*. The use of a back-propagation neural network for plant identification was successfully performed in previous works such as Mancuso and Nicese (1999) in olive, Mancuso *et al*. (1999) in chestnut, Mariño and Tressens (2001) in *Rollinia*, Mancuso (2002) in grapevine, Clark (2004) in *Tilia* and Mugnai *et al*. (2008) in *Camellia japonica*. This method has a good recognition performance and can be effectively used to differentiate *Banksia* species successfully through the use of phyllometric parameters, as almost all the tested accessions were well identified and discriminated by the network. Moreover, BPNN has the advantage of being able to resolve subtle associations between characters, and of making incomplete data (i.e. absence of *Banksia* flower parameters such as styles colour or dimensions) useful in species diagnostics.

Our BPNN appeared to be a powerful tool for discriminating among different *Banksia* ecotypes or forms belonging to the same species. For example, discrimination of the different *B. brownii* forms, for which there is not yet formal recognition of varieties or cultivars, was successful. Two of these forms were easily distinguishable; *B. brownii* 'Mountain' (from the Stirling Ranges, the typical location of this species) had leaves with a more leathery consistency than those of *B. brownii* 'Tree', which, as the name implies, has a tree-like growth form. The third accession, *B. brownii* 'Intermediate', is difficult to distinguish from the 'Tree' form and was included in this analysis because of a slight difference in habit (more bushy) and flowering time. Analysis of output values and %Conf led to a good discrimination among two accessions, *B. brownii* 'Tree' and *B. brownii* 'Mountain', whereas *B. brownii* 'Intermediate' confirmed its morphological closeness to the other two forms, with the presence of similarity peaks in the output values.

In our case, the creation of an ANN based on morphological and digital features of leaves, by using a large number (40) scanned images *per* specimen, can lead to the effective recognition of *Banksia* morphotypes, even though particular care must be put in the initial choice and collection of the plant material, which must be healthy and well developed and free of growth anomalies. Moreover, as our BPNN needs a complete set of data, all characters measured in the training set need to be measured also on the specimen to be identified to increase the ANN identification-process performance. Although collection of multiple samples is time-consuming, it is a necessary part of developing an effective ANN. In ecological and botanical studies it is generally important to be able to recognise species *in situ*, and specimens with flowers are not always available. Facing this problem, a network exclusively based on data obtained from digital images of the leaves can be effectively and successfully used, although not always unequivocally, to discriminate among *Banksia* accessions.

Despite the powerful discriminating capacity of the BPNN on the basis of phyllometric parameters, some limitations occur. They are largely the same as those of a human expert, namely that success depends on the quantity, validity and accuracy of training data. Accurate identification of specimens in the training sets is an essential prerequisite to the good functioning of any automated identification system. Therefore, the data used for the training phase must be representative of the situation to be modelled. It is also important to assure equal sizes of training sets in order to prevent error rates for the smaller category and a lower accuracy of identification (Al-Haddad *et al*. 2000). Nevertheless, the speed, reliability and effectiveness of the approach used here provide an indication that ANN may be a useful adjunct for supporting more traditional taxonomic approaches. Most studies of automated identification systems have employed training sets with relatively small numbers of specimens (i.e. 5–10) per species. However, larger sets would be required to distinguish effectively between species that are narrowly separated (Gaston and O'Neill 2004), especially where any marked variation in the size of training sets for different species may reduce the accuracy of identifications. In fact, it is well known that ANNs train well and learn to generalise best when presented with data rich in variation (Clark 2004) to cover intraspecific biological variation, although some difficulties can also be encountered. For example, individuals of a given species may vary in their morphology, and closely related species may be extremely similar to one another. So, ideally, an automated species-identification system needs not only to be able to match an individual specimen with one of a set of known species, but if necessary it should also be able to refuse it as belonging to a species that is not part of this set. The ability to recognise unknowns is also essential, because when natural samples are analysed it is likely that several species that have not been used for training the network will be encountered. In this case, the ANN would tend to identify specimens of other species as belonging to one of the training set, leading to a false identification. This problem can partly be overcome by adjusting the post-processing phase of ANN modelling, by setting accept and reject thresholds for the output activation values as confidence limits. Only if the output value for a corresponding specimen (accession) exceeded the level of acceptance and all other accessions concurrently failed to exceed the reject threshold, the identification would be made. Otherwise, the identity of the specimen would remain unknown.

To produce a complete and totally effective ANN identification system for *Banksia*, capable of handling the genus throughout the range of all taxa, it would be important to insert in the ANN construction process other peculiar morphological parameters, e.g. *Banksia* flower characters such as style-hook dimension and shape. The inclusion of the

morphometric parameters of the flower should lead to a more powerful, detailed and informative network, with a complete capacity to discriminate. This approach should also be useful and helpful in improving the capability of the current ANN in discriminating among all *Banksia* species throughout the range of all taxa; however, it is not always easy to obtain morphological parameters of the flower throughout the year.

In conclusion, the application of a BPNN is proposed as a complementary method of botanical identification, being capable of separating a wide and comprehensive array of *Banksia* accessions on the basis of leaf morphological characters alone. The main characteristics of this method of data treatment for discrimination purposes are (1) its usefulness, taking into account that an appropriate neural-network architecture must be preliminarily chosen on the basis of the number of specimens and the kinds of parameters for identification, (2) its improved speed of data analysis compared with the traditional methods for plant identification (i.e. dichotomous keys), (3) its easiness of use, with no specialist skills (i.e. molecular approaches) being necessary, and (4) the low cost for identification process, as analysis can be performed with a stock-standard flat-bed scanner connected to a PC, with analyses undertaken by using freeware software for morphological characterisation and the construction of the ANN.

Acknowledgements

The authors thank Kings Park and Botanic Garden staff (Dr Matthew Barrett, Russell Barrett, Mr Bob Dixon) and the institution for hosting this project (providing plant material, scanner, travelling expenses) and for assistance with the identification of *Banksia*; Mr and Mrs Collins, owners of 'The Banksia Farm', for access to their collection of *Banksia* species from eastern Australia.

References

Al-Haddad L, Morris CW, Boddy L (2000) Training radial basis function neural networks: effects of training set size and imbalanced training sets. *Journal of Microbiological Methods* **43**, 33–44. doi: 10.1016/S0167-7012(00)00202-5

Balfoort HW, Snoek J, Smits JRM, Breedveld LW, Hofstraat JW, Ringelberg J (1992) Automatic identification of algae: neural network analysis of flow cytometric data. *Journal of Plankton Research* **14**, 575–589. doi: 10.1093/plankt/14.4.575

Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, Burkill PH (2000) Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. *Marine Ecology Progress Series* **195**, 47–59. doi: 10.3354/meps195047

Chesmore D (2000) Methodologies for automating the identification of species. In 'Proceedings of inaugural meeting of the BioNET-INTERNATIONAL group for computer-aided taxonomy (BIGCAT)'. (Eds D Chesmore, L Yorke, P Bridge, S Gallagher) pp. 3–12. BioNET-INTERNATIONAL Monthly Bulletin. (BioNET-INTERNATIONAL: London)

Chesmore D (2007) The automated identification of taxa: concepts and applications. In 'Automated taxon identification in systematics: theory, approaches and applications'. (Ed. N MacLeod) pp. 83–100. (CRC Press: Boca Raton, FL)

Chesmore D, Ohya E (2004) Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. *Bulletin of Entomological Research* **94**, 319–330. doi: 10.1079/BER2004306

Clark JY (2003) Artificial neural network for species identification by taxonomists. *Biosystems* **72**, 131–147. doi: 10.1016/S0303-2647(03)00139-4

Clark JY (2004) Identification of botanical specimens using artificial neural networks. In 'Proceedings of the 2004 IEEE symposium on computational intelligence in bioinformatics and computational biology (CIBCB)'. pp. 87–94. (CIBCB: La Jolla, CA)

Clark JY, Warwick K (1998) Artificial keys for botanical identification using a multilayer perceptron neural network (MLP). *Artificial Intelligence Review* **12**, 95–115. doi: 10.1023/A:1006544506273

Do MT, Harp JM, Norris KC (1999) A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research* **89**, 217–224. doi: 10.1017/S0007485399000334

Du JX, Huang DS, Wang XF, Gu X (2007) Shape recognition based on neural networks trained by differential evolution algorithm. *Neurocomputing* **70**, 896–903.

Gaston KJ, O'Neill MA (2004) Automated species identification: why not? *Philosophical Transactions of the Royal Society of London. B* **359**, 655–667. doi: 10.1098/rstb.2003.1442

George AS (1981) The genus *Banksia* L.f. (Proteaceae). *Nuytsia* **3**, 239–473.

George AS (1988) New taxa and notes on *Banksia* L.f. (Proteaceae). *Nuytsia* **6**, 309–317.

George AS (1999) *Banksia*. In 'Flora of Australia. Vol. 17B, Proteaceae 3. *Hakea* to *Dryandra*'. (Ed. A Wilson) pp. 175–251. (ABRS/CSIRO: Melbourne)

Giacomini M, Ruggiero C, Calegari L, Bertone S (2000) Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data. *Journal of Microbiological Methods* **43**, 45–54. doi: 10.1016/S0167-7012(00)00203-7

Ginoris YP, Amaral AL, Nicolau A, Ferreira EC, Coelho MAZ (2007) Recognition of protozoa and metazoa using image analysis tools, discriminant analysis, neural networks and decision trees. *Analytica Chimica Acta* **595**, 160–169. doi: 10.1016/j.aca.2006.12.055

Hernández-Borges J, Corbella-Tena R, Rodriguez-Delgado MA, Garcia-Montelongo FJ, Havel J (2004) Content of aliphatic hydrocarbons in limpets as a new way for classification of species using artificial neural networks. *Chemosphere* **54**, 1059–1069. doi: 10.1016/j.chemosphere.2003.09.042

Mancuso S (2002) Discrimination of grapevine (*Vitis vinifera* L.) leaf shape by fractal spectrum. *Vitis* **41**, 137–142.

Mancuso S, Nicese FP (1999) Identifying olive (*Olea europaea* L.) cultivars using artificial neural networks. *Journal of the American Society for Horticultural Science* **124**, 527–531.

Mancuso S, Ferrini F, Nicese FP (1999) Chestnut (*Castanea sativa* L.) genotype identification: an artificial neural network approach. *Journal of Horticultural Science & Biotechnology* **74**, 777–784.

Mariño SI, Tressens SG (2001) Artificial neural networks application in the identification of three species of Rollinia (Annonaceae). *Annales Botanici Fennici* **38**, 215–224.

Mast AR, Givnish TJ (2002) Historical biogeography and the origin of stomatal distributions in *Banksia* and *Dryandra* (Proteaceae) based on their cpDNA phylogeny. *American Journal of Botany* **89**, 1311–1323. doi: 10.3732/ajb.89.8.1311

Mast AR, Thiele K (2007) The transfer of *Dryandra* R.Br. to *Banksia* L.f. (Proteaceae). *Australian Systematic Botany* **20**, 63–71. doi: 10.1071/SB06016

Mast AR, Jones EH, Havery SP (2005) An assessment of old and new DNA sequence evidence for the paraphyly of *Banksia* with respect to *Dryandra* (Proteaceae). *Australian Systematic Botany* **18**, 75–88. doi: 10.1071/SB04015

Morgan A, Boddy L, Mordue JEM, Morris CW (1998) Evaluation of artificial neural networks for fungal identification, employing morphometric data from spores of *Pestalotiopsis* species. *Mycological Research* **102**, 975–984. doi: 10.1017/S0953756297005947

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

Mouwen DJM, Capita R, Alonso-Calleja C, Prieto-Gómez J, Prieto M (2006) Artificial neural network based identification of *Campylobacter* species by Fourier transform infrared spectroscopy. *Journal of Microbiological Methods* **67**, 131–140. doi: 10.1016/j.mimet.2006.03.012

Mugnai S, Pandolfi C, Azzarello E, Masi E, Mancuso S (2008) *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. *Plant Systematics and Evolution* **270**, 95–108. doi: 10.1007/s00606-007-0601-7

O'Neill MA (2007) DAISY: a practical computer-based tool for semi-automated species identification. In 'Automated taxon identification in systematics: theory, approaches and applications'. (Ed. N MacLeod) pp. 101–113. (CRC Press: Boca Raton, FL)

Pandolfi C, Mugnai S, Azzarello E, Masi E, Mancuso S (2006) Fractal geometry and neural networks for the identification and characterization of ornamental plants. In 'Floriculture, ornamental and plant biotechnology: advances and topical issues. Vol. IV'. (Ed. J Teixiera da Silva) pp. 213–225. (Global Science Books: Kyoto)

Sahin N, Aydin S (2006) Identification of oxalotrophic bacteria by neural network analysis of numerical phenetic data. *Folia Microbiologica* **51**, 87–91. doi: 10.1007/BF02932161

Smits JRM, Breedveld LW, Derksen MJW, Kateman G, Balfoort HW, Snoek J, Hofstraat JW (1992) Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. *Analytica Chimica Acta* **258**, 11–25. doi: 10.1016/0003-2670(92)85193-A

Thiele K, Ladiges PY (1994) The *Banksia integrifolia* L.f. species complex (Proteaceae). *Australian Systematic Botany* **7**, 393–408. doi: 10.1071/SB9940393

Thiele K, Ladiges PY (1996) A cladistic analysis of *Banksia* (Proteaceae). *Australian Systematic Botany* **9**, 661–733. doi: 10.1071/SB9940393

Van hara 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. doi: 10.2478/s11756-007-0089-1

Walsh S, MacLeod N, O'Neill MA (2007) O spot the penguin: can reliable taxonomic identifications be made using isolated foot bones? In 'Automated taxon identification in systematics: theory, approaches and applications'. (Ed. N MacLeod) pp. 225–238. (CRC Press: Boca Raton, FL)

Weeks PJD, Gaston KJ (1997) Image analysis, neural networks, and the taxonomic impediment to biodiversity studies. *Biodiversity and Conservation* **6**, 263–274. doi: 10.1023/A:1018348204573

Wheeler QD (2007) Digital innovation and taxonomy's finest hour. In 'Automated taxon identification in systematics: theory, approaches and applications'. (Ed. N MacLeod) pp. 18–23. (CRC Press: Boca Raton, FL)

Wilkins MF, Boddy L, Morris CW, Jonker RR (1999) Identification of phytoplankton from flow cytometry data by using radial basis function neural networks. *Applied and Environmental Microbiology* **65**, 4404–4410.

Zurada F, Jacek M (1992) 'Introduction to artificial systems.' (PWS Publishing: Pacific Grove, CA)

Manuscript received 15 January 2008, accepted 19 January 2009