Alleviating the taxonomic impediment in diatoms: prospects for automatic and web-based identification systems

Mann, D.G.[1*], S.J.M. Droop[1], Y.A. Hicks[2], A.D. Marshall[3], R.R. Martin[3] and P.L. Rosin[3]

[1] Royal Botanic Garden, Edinburgh EH3 5LR, Scotland, U.K.

[2] School of Engineering, Cardiff University, Queen's Buildings, P.O. Box 925, Cardiff CF24 0YF, Wales, UK

[3] School of Computer Science, Cardiff University, Queen's Buildings, Newport Road, Cardiff CF24 3AA, Wales, UK.

* email: d.mann@rbge.org.uk

Mann, D.G, S.J.M. Droop, Y.A. Hicks, A.D. Marshall, R.R. Martin and P.L. Rosin (2005). Alleviating the taxonomic impediment in diatoms: prospects for automatic and web-based identification systems.

Abstract: Automated identification of diatoms can be based on morphology, a combination of shape, size and physiological characteristics (in flow cytometry), or genetic markers. For palaeoecological, stratigraphic and many ecological studies, however, only the first of these – morphology – is available. We review three prototype systems that have been developed and tested since 1970, that use the morphology of the frustule as a basis for automated identification: the optical diffractometry approach of the Cairns–Almeida group, and the ADIAC and DIADIST projects. The latter two extract features from digital micrographs and use these for identification, either via a decision tree or by distance measures. DIADIST also models variation and, unlike ADIAC, its algorithms can be used to characterize drawn diatoms as well as photographs, allowing cross reference (visual indexing) between different types of published and unpublished image. Furthermore, DIADIST's algorithms can synthesise new images. ADIAC and DIADIST both achieve correct identification rates of > 95% in tests. The taxonomic impediment to diatom research and the use of diatoms in ecological monitoring can also be lessened by imaginative use of the World-Wide Web, e.g. to disseminate images of type and authenticated material as ‘focusable’ image stacks, and by making rare literature available both as scanned copy and as iconographs of published images.

In the last 20 years there has been a remarkable increase in interest in diatoms, because of their importance in carbon and nutrient cycling (e.g. see Mann 1999), their use in palaeoecological reconstruction and water quality monitoring (Stoermer & Smol 1999), and their ability to produce patterned silica structures at physiological temperatures and pressures (e.g. Pickett-Heaps et al. 1990, Kröger et al. 1999, 2002). As with many other kinds of organisms, however, increased interest in the biology and ecology of the group has not translated into increased support for basic taxonomy. Valuable taxonomic work continues to be done, but the total output is unimpressive in the context of the number of species already known and the likelihood that this represents only a small fraction of those that actually exist (perhaps 200,000 species: Mann & Droop 1996, Mann 1999, Behnke et al. 2004, Mann et al. 2004). There have been attempts to provide a higher-level taxonomic framework for diatoms (e.g. Round et al. 1990), a freshwater diatom flora has been produced for Europe (Krammer & Lange-Bertalot 1986, 1988, 1991a, b), and there are detailed treatments of particular genera (e.g. Krammer 1992), but their inadequacy is only too apparent from the speed with which they have become outdated (e.g. Medlin & Kaczmarska 2004, Lange-Bertalot & Metzeltin 1996, Krammer 2000). And, as often happens, recent changes in taxonomy have made existing identification aids (e.g. dichotomous keys in the major floras) largely obsolete, and there were few enough of these anyway! Ideally, more taxonomists should be employed, to provide a sound, easily accessible and intelligible taxonomy for other scientists to use. This appears to be a forlorn hope. So, instead, we must seek ways to make more effective use of the limited expertise and information that are available. In this paper, we review relevant progress in computer-aided identification, updating the report by Bayer et al. (2001), and in use of the World-Wide Web (WWW) in diatom taxonomy.

General considerations

There have been many attempts to automate the identification of biological material. The basic reason is simple: accurate, consistent identification by human beings is expensive. Each taxonomically competent scientist needs individual training, resources (e.g. a reference library), a reasonable working environment, a salary (perhaps not as much as he or she would like), and time off to sleep and have a holiday. Furthermore, even when a species is recognized immediately, without recourse to texts, recording the identification takes a significant amount of time. Identifying and counting many hundreds of organisms per day, as is necessary in many palaeoecological, stratigraphic or ecological studies, is tedious and exhausting, and errors are easily made in transcribing results. Ideally, therefore, we should supplement or replace people with

machines that do not tire, sleep or take breaks, work consistently and with defined accuracy, and record data faithfully.

The characters used for identifying taxa need not be the same as those used for classifying them. For example, geographical distribution is not an intrinsic feature of an organism but something that may have been influenced by historical or anthropological factors. It is therefore inappropriate to use it as a character during classification. Nevertheless, it is entirely legitimate to take geography into account when developing *identification* systems. In Britain, there are no native tigers and so 'mammal with sharp canines in the upper and lower jaw and striped fur' will almost always uniquely identify Scottish wild cats (*Felis silvestris grampia* Ragni & Randi) and some feral domestic cats. It is only if illustrations show that the animal in question is manifestly *not* a Scottish wild cat that it becomes necessary survey the remainder of the world's mammals and conclude, perhaps, that a pet tiger or lynx has escaped from captivity. Also in contrast to classification, identification systems can legitimately use the same information in several different ways, deliberately make no distinction between homology with homoplasy, and break up a continuously varying character into quite arbitrary character states, if it helps to pin the right name on a specimen.

Identification can in theory be based on any features that vary among organisms or parts of organisms, and automated identification is no different. In practice, however, most automated methods depend on use of morphology *sensu lato* (size, shape and pattern) or on a restricted range of biochemical or genetic properties, e.g. short DNA sequences. A special case, useful in species-poor communities of microorganisms such as phytoplankton, is the use of biochemical and spectral properties, size and shape in combination to allow rapid enumeration via flow cytometry (see also Gaston & O'Neill 2004)..

Recently, there have been proposals for kingdom-wide 'bar-coding' of species via the nucleotide sequences of appropriately variable genes (Hebert et al. 2003a, b, Blaxter 2004). Although controversial as an overall replacement for orthodox, morphology-based taxonomy (e.g. Scotland et al. 2003, Will & Rubinoff 2004), in more limited circumstances, for identification within particular well-studied but difficult groups, DNA bar-coding may be the best way forward. A good candidate for such approaches in diatoms is *Pseudo-nitzschia*, where identification on the basis of structure and shape is tedious and often uncertain. Even those who work intensively with *Pseudo-nitzschia* can make identifications that later prove incorrect, such as the '*P. pseudodelicatissima*' clones studied by Davidovich & Bates (1998), which fall into two groups, one corresponding to *P. calliantha*, the other to either *P. pseudodelicatissima* or *P.*

cuspidata (Lundholm et al. 2003). DNA approaches have already been used for routine identification in *Pseudo-nitzschia* and with increasingly good sampling of genetic variation there is every prospect of a fully automated, comprehensive system for the genus. For truly cryptic species, use of DNA methods is often now the way they are detected (e.g. Blaxter et al. 2004 for tardigrades) and it is the only possible basis for automated identification. DNA methods could in theory be used for any group of diatoms, but without massive effort to grow species in culture for DNA extraction and 'calibration' by taxonomists, it will be impossible to relate any new DNA-based classification and identification system to previous data tied to the existing (admittedly imperfect) Linnean nomenclature.

In future, where material can be collected alive or fixed to preserve DNA, most automated systems will probably be DNA-based, because the technology is relatively simple, well established and increasingly inexpensive, and needs no special development for particular groups or organisms, other than the selection of appropriate genetic markers. By contrast, automated identification systems that use morphology are always challenging to construct and work for relatively small sets of objects, because of the huge variation in Bauplan among living and fossil organisms. Despite this difficulty, however, there is a strong case for developing morphology-based identification systems for diatoms, because DNA approaches cannot be used for the identification of frustules, which are the essential basis for palaeoecological, stratigraphic, and many ecological studies. Other organisms where similar considerations apply include foraminifera, coccolithophorids and pollen and for all of these there have been attempts to develop automated identification systems using shape and pattern (Beaufort & Dollfus 2004, Bollmann et al. 2004, France et al. 1997, 2000, 2004, Yu et al. 1996).

The special problems posed by diatoms

Diatom frustules have some properties that make them particularly suitable for development of automated identification methods. They are generally rigid structures and so do not collapse or shrink. Because of their box-like construction, many frustules have only one or a few likely orientations after preparation for microscopy and in each orientation they exhibit relatively simple outline shapes. The pattern of markings on the valves has one to a few dominant frequencies, which remain fairly constant during the life cycle.

Diatoms also have 'undesirable' properties. They are small, with many features that are dimensionally close to the wavelengths of visible light, so that interpretation using the light microscope is greatly complicated by diffraction. Diffraction and the low depth of

focus of high resolution objectives also make separation of a clean diatom contour from the background difficult. Diatoms also have few features that can be used for landmarking, so that a whole range of morphometric (and hence identification) tools are unavailable (cf. Bookstein 1991). Perhaps worst, shape and size and pattern change during the life cycle, so that small frustules of one species may resemble large frustules of the same species less than they do small frustules of a different species (cf. Hustedt 1937). There are two ways to cope with this variation in an automated system. One is simply to include many examples of each species in the set used to train the system, so that any new, unidentified specimen is likely to find a counterpart in the training set. The other is to attempt to model the variation.

Early attempts at automated identification of diatoms

The first attempt to use computers to automate identification of diatoms was remarkably early, in the 1970s. Motivated in part by the need for large sample counts in the environmental quality monitoring schemes developed by Patrick and colleagues (e.g. Patrick et al. 1954, Patrick 1967, Cairns et al. 1970), Cairns, Almeida and their co-workers made a prototype system for automated identification, using optical diffractometry. The effort put into this was considerable (Almeida et al. 1971, 1977, 1979, Almeida & Eu 1975, Almeida & Fujii 1979, Cairns et al. 1972, 1973, 1974a, b, 1977a, b, 1979, 1982, Case et al. 1978, Dickson et al. 1976, Fujii & Almeida 1979, Fujii et al. 1980, Partin et al. 1979), but no further development appears to have occurred after the early 1980s. By the time Cairns et al. (1982) summarized progress, some of the problems inherent to their earlier work – namely, use of 35 mm transparencies as source material, rather than input directly from the microscope, and sensitivity of the method to orientation of the diatoms – had been overcome. Others, such as the sensitivity of their method to variation in appearance of diatoms during life cycle changes, were to be addressed by use of multiple exemplars per species, as in more recent approaches (see below). However, the method was not developed further after the initial funding, probably because a series of holographic filters would have to be developed for each species and used in all orientations in highly specialized equipment. Nevertheless, the basic principle underlying the Cairns–Almeida system – applying Fourier transforms to detect and characterize regularities of outline and pattern – is certainly appropriate for diatoms. Fourier analysis of outline shape and texture forms part of the more recent ADIAC and DIADIST initiatives.

Image analysis and ADIAC

Bayer et al. (2001) and du Buf & Bayer (2002a) have described how image analysis and multivariate statistics were introduced to diatom taxonomy in the 1980s and 1990s. This work, principally by Stoermer's group at Michigan, introduced new shape descriptors (descriptions had previously been almost entirely verbal), particularly Legendre polynomials and Fourier coefficients (e.g. Stoermer & Ladewski 1982, Mou & Stoermer 1992). Further papers by this group have been published recently (Pappas et al. 2001, Rhode et al. 2001, Pappas & Stoermer 2003), while other researchers have used simpler characterizations of shape or pattern (e.g. Droop 1994, Droop et al. 2000). The purpose throughout has been to develop and apply morphometric methods in the investigation of variation in groups where the species boundaries are unclear; the methods are used to discover groups that may be separate species. Mann et al. (2004) tested several morphometric methods for their ability to separate subtly different species in the *Sellaphora pupula* complex. As noted by du Buf & Bayer (2002b), 'all of [these] .. studies show that the extraction of outline/shape features, combined with multivariate analysis methods such as principal component analysis, is a very powerful tool for resolving the subtle morphological variation in diatoms, at the level of species and beyond.' This is also an important pre-requisite for automating identification.

The ADIAC project ran from 1998 to 2001, leading to several papers (Bayer et al. 2001, Ciobanu et al. 2000, du Buf et al. 2000, Fischer & Bunke 2001, 2002a, Fischer et al. 2000a, b, 2002, Loke et al. 2002, 2004, Wilkinson et al. 2000) and a book summarizing the principal results (du Buf & Bayer 2002). Since the completion of the book, further papers have been published, either extending the ADIAC work itself (Ambauen et al. 2003, Jalba & Roerdink 2003, Jalba et al. 2004) or using the ADIAC image database (http://rbg-web2.rbge.org.uk/ADIAC/pubdat/downloads/public_images.htm) to develop and test further shape descriptors (Rosin 2003, 2004, Zunic & Rosin 2004). There is related work by Forero-Vargas et al. (2003).

In ADIAC, all of the shape and texture measures that had previously been tried on diatoms (e.g. Legendre polynomials, Fourier analysis, rectangularity, size and aspect ratio, stria density) were used, together with several others that had either been ignored previously for diatoms (measures of symmetry, global shape, triangularity, circularity, moment invariants, grey-level co-occurrence, use of Gabor filters, curvature scale space, etc) or were developed specifically (e.g. stria orientation and the point where stria orientation changes from radial to convergent, axial area width, contour segment analysis, contour profiling, etc) (Fischer & Bunke 2002b, Loke & du Buf 2002, Ciobanu & du Buf, Santos & du Buf 2002, Wilkinson et al. 2002). Many of these (Table 1) are different ways of quantifying the same class of features, e.g. shape or pole shape or

striation pattern, but it is difficult *a priori* to decide which way of 'viewing' the data will provide the best way to discriminate between two taxa. In ADIAC, evaluation of particular descriptors was made *a posteriori*, by seeing how well they performed in correctly identifying a test set of images, after the system had been supplied with a training set containing the same taxa. A decision tree algorithm (e.g. the C4.5 algorithm) was generally used as the classifier, which examines the training set derived from known specimens and finds which features will best split the set into smaller subsets and ultimately (if possible) into groups comprising single taxa. If the training set is changed or augmented, the algorithm produces a new decision tree. The system is explicit, showing which decisions are made at each step, and the performance of particular characters can be monitored.

Westenberg & Roerdink (2002) summarized the performance of different descriptors in identifying 37 pennate diatom species as diverse as *Cocconeis placentula*, *Navicula radiosa*, *Gyrosigma acuminatum*, *Gomphonema augur*, *Epithemia sorex*, *Nitzschia hantzschiana*, *Surirella brebissonii* and *Tabellaria flocculosa*. Particular sets of characters performed quite well on their own; Fourier descriptors of outline shape, for example, gave 84% correct identification. Identification rates were much higher, however, when feature sets were used in combination. Thus, when all contour features (Table 1) were used together, rather than Fourier descriptors alone, identification rates rose to 92%, and use of all features (over 300, both shape and pattern) gave the best rate, of nearly 97%.

Kelly et al. (2002) compared ADIAC performance with human identifications. This is not easy. Unlike the computer, which, for the mixed pennate diatom set, only 'knew' about 37 species, diatomists will be aware of many hundreds or thousands of other taxa that are candidate identifications. The low average 'success' achieved by trained diatomists (63%) in identifying members of the mixed pennate data-set is therefore difficult to interpret, though alarming for those planning ecological or palaeoecological surveillance. More instructive was a test in which diatomists were supplied with a training set of images and descriptions of six demes of *Sellaphora pupula* (all very similar morphologically, none of them yet treated in published floras and papers), to help them identify the demes correctly. All of them therefore started with an authoritative guide and knew that only those six demes were present in the test set. Here, the 10 scores varied from 60 to 98.3% (Kelly et al. 2002). The best scores were achieved, not surprisingly, by two people who had had extensive research experience of *S. pupula*. Overall, the average achieved by humans on the *Sellaphora* set was 82%.

Legendre polynomial descriptors and contour profiling, however, gave close to 100% success (Ciobanu & du Buf 2002).

In a very recent paper (developing the ADIAC work summarized by Wilkinson et al. 2002), in which they used hat-transform methods to characterize different levels of morphological detail of the outline and 'interior' pattern, Jalba et al. (2004) have reported success rates of 99.6%. Clearly, therefore, automatic identification of diatoms is feasible.

All the test specimens used in identification tests were carefully selected and input as digital images, all in the same orientation. All were clean and entire, with no significant overlapping material. Most strewn diatom slides do not provide such ideal material. If samples are diluted before being placed on cover-slips, overlaps can be minimized, but then search times are much longer. Automatic slide scanning is therefore important. During ADIAC, progress was made on developing automatic slide scanning (Pech-Pacheco & Cristóbal 2002), but this remains the Achilles heel of the ADIAC system, together with the significant computing time needed per specimen (which could be overcome by parallel processing, perhaps using GRID technology).

DIADIST

The ADIAC methods were designed to operate on grey-scale digital images taken using the light microscope. However, there are other classes of image that have traditionally been important in diatom taxonomy, viz. drawings and printed (screened) photographs. There are huge numbers of these. The Fritsch collection of illustrations of freshwater & terrestrial algae at Windermere, U.K., contains c. 500,000 images of microalgae and protists, extracted from over 15000 publications ('Fritsch' at http://www.idc.nl/catalog/), and perhaps 20% of this collection are diatoms. Inclusion of all marine microalgae would expand this massive collection still further. The Fritsch collection images – available also in microfiches – can be searched by the taxon names that have been attached to them (principally by their authors) (Eloranta 1987), or browsed by taxon alphabetically within the major algal groups. Ideally, however, we would not only like to ask questions like 'what illustrations are there of *Sellaphora* species' but also 'has anyone seen a diatom that looks like this?' In other words, we need content-based 'visual indexing' of image databases, not simply textual indexing.

The Diatom and Desmid Identification by Shape and Texture (DIADIST) project [http://rbg-web2.rbge.org.uk/DIADIST/] focused on ways to cross-reference different types of image – photographs and drawings – by determining salient details of pattern and

outline and using a relatively new mathematical approach to model phenotypic and life-cycle variation. After exploration (Hicks et al. 2002) and rejection of pseudo-landmarking as a basis for modelling how shape changes during the life cycle, we decided to characterize shape using Fourier descriptors, which has already proved effective in morphometric investigations (e.g. Mou & Stoermer 1992, Pappas et al. 2001) and in ADIAC (see above). The outline is treated as a cyclic shape that can be approximated by summing component sinusoidal waveforms, each with a characteristic frequency. The amplitude and phase of each component waveform are the Fourier descriptors and we described diatom outlines with a 200 element vector, consisting of 100 amplitude values and 100 corresponding phase angles (Hicks et al. 2004). Striation patterns were also characterized using 2-D Fourier analysis. Unlike in the Cairns–Almeida system, which treated the valve as having a single complex pattern, we analysed variation in the pattern across and along the valve. A fast Fourier transform (FFT) was performed within a square window (of a size sufficient to include at least

three striae) centred on each pixel in turn within the diatom contour, to detect how the frequency, orientation and intensity of pattern varies within the valve (Figs 1, 2). After filtering to remove certain frequencies and orientations of pattern (such as longitudinal striae, which we did not wish to consider in this prototype system), we obtained three maps from FFT processing, giving for each pixel (1) the dominant frequency (Fig. 3), (2) the dominant orientation (Fig. 4).

It is also important for taxonomy to detect areas, such as the sternum of pennate diatoms, where there is no pattern (although there may be unique structures in or near the sternum, such as the raphe, rimoportulae or stigmata). For this, a large square window is inappropriate. We therefore performed a second FFT analysis on each valve, now using an apically elongate window (Fig. 1), to detect areas of very low pattern intensity (low energy values). In most pennate diatoms, this detects the axial and central areas (Fig. 5). Their edges were defined by suitable thresholding (Fig. 6)and we characterized their outline by fitting a cubic spline curve (with 19 control points) to the border of the sternum in each quadrant of the valve. Fourier analysis could be used to describe the shape of the sternum, as with the outline, but the sternum does not have a clearly defined border (because the sternum is continuous with the transapical ribs separating the striae) and it is best to use a measure that applies more smoothing to the sternum outline. To characterize the pattern for classification or identification, we divided each quadrant of the cell into three sectors (polar, median and central: the borders of each sector were equally spaced along the sternum outline), because this

allows satisfactory characterization of stria patterns, which are generally either radial, parallel or convergent in each quadrant. The average orientation and frequency of the stria pattern was then determined for each sector as the average of all orientation and frequency values, weighted by the corresponding energies. This yielded 24 orientation and frequency values.

Overall, then, each valve was characterized by 200 Fourier descriptors of the outline, 76 coordinates of the 38 spline control points, 24 orientation and frequency values for the stria pattern, and contour length, which measures absolute size. If populations of valves of a given species have been sampled, the main trends in the variation pattern (which will usually be dominated by the changes that accompany size reduction during the life cycle) can be modelled for each species. To do this, we reduced the dimensionality of the data by principal component analysis and fitted principal curves (Figs 8, 9) (Hicks et al. 2004: http://www.cs.cf.ac.uk/diadist/articles.htm). Unknown specimens can be identified by extracting the same set of features as have been used for the training set and calculating the Euclidean distance between the coordinates of the specimen in feature space and the nearest principal curves. Alternatively, if no modelling via principal curves were possible because of very small numbers of exemplars (or single images), an image collection could be searched on the basis of inter-specimen distance in feature space, with whatever selectivity or weighting one might wish to impose (e.g. ‘what other images show similarly lanceolate valves, whatever their stria pattern?’).

Tests on the DIADIST methods using a subset of the ADIAC images (178 images representing 13 species) gave a success rate of 96.6% using all characters (shape, texture and contour length). This is similar to the success rate achieved in 2002 by ADIAC (Westenberg & Roerdink 2002), but ADIAC included more species, including some (e.g. *Nitzschia*) that have patterns unsuitable for the current DIADIST approach. Nevertheless, the identification rate is very good and it is easy to see how the windowing approach could be adapted to work with structurally asymmetrical diatoms where there is no obvious sternum (most Bacillariaceae), or with diatoms (e.g. *Lyrella*) that have lateral sterna as well as an axial raphe sternum. There is also nothing magical about the particular sizes and shapes of window we used for the striae and axial areas, and there is no reason why the number of sectors per valve should be 12 as opposed to 24, which could allow stria curvature to be detected (each of the 12 sectors would be divided into an inner and an outer part), or some other number. The FFT also reveals patterns along the striae (pore spacings), which we ignored for simplicity. Such modifications could be highly beneficial in particular cases, and there might therefore

be an initial pre-processing of images to determine which particular algorithms are most appropriate and which subsets of an image collection should be used as comparators. To deal with centric diatoms, a more divergent strategy will be needed.

As mentioned previously, the DIADIST algorithms were designed to work not only on grey-scale photographic images, but also on bitmap images or drawings, and preliminary results using drawings are encouraging (Hicks et al., submitted). Some drawing styles can cause difficulties. Where striae are drawn as single lines or broken lines consisting of closely spaced dots, the Fourier windowing method works well. However, if the striae are wide, as in many *Pinnularia* and some *Navicula* sensu stricto species, diatomists often draw the boundaries of the striae (like an elongate sausage), rather than filling them in. Here, Fourier analysis can 'regard' both lines as equivalent elements of pattern, so that the stria frequency is estimated to be double its real value. Such problems may or may not apply to photographic images of the same species.

Errors are easily detected because all of the DIADIST shape and texture measures are readily 'invertible', in the sense that, given the numerical shape and texture descriptors, a new summary image ('drawing') can be synthesized for comparison with the originals (Fig. 7, compare Fig. 1), or as an independent record of morphology. By contrast, some shape and pattern descriptors do not allow useful reconstruction of the original: for example, many different shapes can have the same rectangularity. In the DIADIST methods, the outline shape can be reconstructed through inverse Fourier transform and adjusted to the correct overall size. Likewise, the cubic spline points determine the shape of the axial and central areas. In order to synthesize a realistic stria pattern, we tried to approximate the natural ontogeny of the valve (e.g. Round et al. 1990, Pickett-Heaps et al. 1990). We therefore grew pattern outwards from the sternum in each of the 12 sectors, with the frequency and orientation appropriate to each, as specified by the pattern descriptors. If striae radiated and became too far apart, another stria was inserted to maintain roughly equal spacing; conversely, when striae converged, one was suppressed. If this is done for a particular specimen, the outcome is usually a reasonable approximation to the original (Figs 1, 7), though it is in no sense a tracing. Where a population has been modelled using principal curves, it is possible to synthesize drawings of virtual specimens that express the principal trends of variation (Fig. 10). Some of these drawings may have no counterpart in the original training set, e.g. because of subdivision of the population into size classes (cf. the populations of *Nitzschia sigmoidea* studied by Mann 1988; see also Nipkow 1927).

Prospects for automated identification and visual indexing

We have reviewed three attempts to alleviate the taxonomic impediment in diatoms that use computer-based image analysis and pattern recognition methods to minimize human involvement in the many tedious processes involved in identifying and counting diatoms. Each initiative can reasonably claim to have established 'proof of concept'. On the other hand, none has produced a working system that can be marketed to scientists and monitoring agencies. Gaston & O'Neill (2004) note that automated systems for other groups of organisms have reached a comparable stage before being essentially abandoned. A problem that always faces projects after the initial exploratory phase is that funding is supposed to be taken over by those who are supposed to stand to gain financially by further development, or by venture capitalists working speculatively on their behalf. Despite the considerable importance of diatoms in monitoring water quality, there is not yet a huge market for diatom identification skills, and the market that does exist is widely dispersed and unequally developed. Hence, we do not anticipate rapid development of the ADIAC and DIADIST approaches. Cairns (2002) suggests that 'a major effort should be initiated to convince policymakers of the value of various types of biological monitoring' and this would certainly provide a better context for further development of automated approaches.

Meanwhile, there is at least one reason to be optimistic: diatoms are becoming model systems for the development of shape descriptors and pattern recognition. There are relatively few good data-sets available for computer vision specialists to experiment upon and the ADIAC images constitute one of them; they have already been used by several groups outside the ADIAC partnership (e.g. Rosin 2003, Zunic & Rosin 2004). This academic interest needs to be encouraged, because it brings us closer to the prospect of a robust commercial or semi-commercial system.

Web-based dissemination of taxonomic information

We conclude with a short section on the use of the World-Wide Web to improve diatom taxonomy. Several important resources are already available, such as the Hustedt collection database (http://www.awi-bremerhaven.de/Research/hustedt1.html), the Academy of Natural Sciences of Philadelphia (ANSP) image database (http://diatom.acnatsci.org/AlgaeImage/), and various materials, particularly the database of diatom genus names, at the California Academy of Science (http://www.calacademy.org/research/diatoms/types/).

One of the worst problems for anyone identifying microalgae is that relevant information is scattered through many books and papers, many of them old and rare. Even in Europe, where there has been active taxonomic research on microalgae since

the early nineteenth century, most phycologists do not have access to a comprehensive library. The obvious solution is to make literature freely available via the WWW, where copyright problems do not exist or can be overcome. Scanned material can be offered as downloadable .pdf files. With more effort, the original material can be reorganized into databases, like the collections of 19th century and early 20th century diatom and desmid illustrations on the DIADIST site (http://rbg-web2.rbge.org.uk/DIADIST/), and we hope to use these to demonstrate visual indexing by the DIADIST algorithms.

Collections of modern photographic images are available on several sites, including the ADIAC collection (http://rbg-web2.rbge.org.uk/ADIAC/db/adiacdb.htm; see also links to other collections at http://rbg-web2.rbge.org.uk/DIADIST/links.htm). Generally, these WWW collections are simply the cyber-equivalents of Schmidt's Atlas (Schmidt et al. 1874–1959). However, new media offer scope for developing novel ways to present images. Type specimens are the ultimate reference material for taxonomy and they are correspondingly precious. Not surprisingly, therefore, some herbaria are unwilling or unable to lend type material and visiting these herbaria is impractical or unaffordable for most people. In fact, however, there is little need for types to have to travel, because all relevant information visible with the light microscope can be captured in a stack of optical sections, which can be made available via the Internet as a movie or as an apparently focusable image. Examples are presented at http://rbg-web2.rbge.org.uk/algae/research/types/types.htm. They can help ensure that names are correctly applied and that vital taxonomic information is accessible to all, at minimal cost. We can therefore democratize taxonomy in a way that has hitherto been impossible and at the same time help to ensure that names are applied more consistently and appropriately, solving many of the problems identified by Mann (1998).

Another old problem that the Web can solve is the expense of providing multiple images of a single taxon to show variation. The only workers who have consistently illustrated many specimens of each taxon are Fukushima, Ko-Bayashi and their coworkers (e.g. Fukushima et al. 2001, 2004), but such an approach is not generally sustainable via printed material. Most journals simply refuse to publish more than a select few representative photographs. By contrast, using the Web, we have been able to make available images of all 684 valves of *Diploneis* demes used by Droop et al. (2000) and the 383 valves of six *Sellaphora* species examined by Mann et al. (2004). A further project we strongly advocate is establishment of a 'virtual herbarium' – a collection of virtual specimens available freely to all and open to contributions from any registered phycologist. The ANSP image database (http://diatom.acnatsci.org/AlgaeImage/) takes a useful step in this direction by allowing comments to be made about particular images.

Ideally, searching the virtual herbarium would be not only using the taxon names attached to the images and associated data, as in a conventional herbarium, but also on the basis of content, using methods like those developed during DIADIST and ADIAC. To avoid the problem that the host of such a herbarium, or some self-appointed committee, might be regarded as having inappropriate and undemocratic control over the herbarium (and possibly, therefore, over diatom taxonomy), the taxonomic framework for the virtual herbarium would be established on 'Prometheus' principles (http://www.dcs.napier.ac.uk/~prometheus/). This allows multiple taxonomies to co-exist in the same database, so that a single specimen can have multiple identities, according to the views of different taxonomists, or even of the same taxonomist at different times. As well as virtual specimens, such a database could hold information on ecological, physiological, reproductive or other characteristics, providing these can be associated with specimens. The specimens themselves could be photographs of diatoms, or published or unpublished drawings (thus allowing the incorporation of autograph material, such as the drawings of Ehrenberg, Grunow etc: see e.g. Lazarus 1998, Jahn et al. 2004).

Finally, there are several types of software, e.g. Lucid (http://www.lucidcentral.com/), Linnaeus (http://www.eti.uva.nl/) or Pankey (http://www.exetersoftware.com/cat/pankey/pankey.html), that allow interactive identification, based on a 'table of attributes' for taxa. This is the simplest type of computerized identification system, corresponding conceptually to the multiaccess and dichotomous keys provided in conventional floras, though with much more flexible access and convenience. They can be made available as stand-alone systems, on CD-ROM, or via the Web. Currently, a Lucid system is being developed for diatoms in UK rivers. The disadvantages of computerized key systems are the labour in constructing the table of attributes and the difficulty of keeping this up-to-date when taxon concepts change. By contrast, systems such as ADIAC and DIADIST are specimen-based, so that if particular specimens are reclassified, the classifier is automatically updated. The advantage of computerized key systems is that they are operable by anyone who has experience of conventional floras and monographs.

Acknowledgements

DIADIST research was supported by BBSRC/EPSRC grants BIO14261 and BIO14262. ADIAC was funded by European Marine Science and Technology programme, contract MAS3-CT97-0122. David Mann thanks the Royal Society for an equipment grant enabling purchase of a Reichert photomicroscope.

References

ALMEIDA, S.P. & J.K.-T. EU (1976): Water pollution monitoring using matched spatial filters. – Appl. Optics **15**: 510–515.

ALMEIDA, S.P. & H. FUJII (1979): Fourier transform differences and averaged similarities in diatoms. – Appl. Optics **18**: 1663–1667.

ALMEIDA, S.P., D. DEL BALZO, J. CAIRNS JR, K.L. DICKSON & G.R. LANZA (1971): Holographic microscopy of diatoms. – Trans. Kansas Acad. Sci. **74**: 257–260.

ALMEIDA, S.P., J.K.T. EU, P.F. LAI, J. CAIRNS JR & K.L. DICKSON (1977): A real-time optical processor for pattern recognition of biological specimens. – In: MAROM, E., A.A. FRIESEM & E. WIENER-AVNEAR (eds): Applications of holography and optical data processing: 573–570. Pergamon Press, Oxford.

ALMEIDA, S.P., S.K. CASE & W.J. DALLAS (1979): Multispectral size-averaged incoherent spatial filtering. – Appl. Optics **18**: 4025–4029.

AMBAUEN R., S. FISCHER & BUNKE H (2003): Graph edit distance with node splitting and merging, and its application to diatom identification. – Lecture Notes Computer Sci. **2726**: 95–106.

BAYER, M.M., M.R. PULLAN, D.G. MANN, S. JUGGINS, A. CIOBANU, L. SANTOS, H. SHAHBAZKIA, H. DU BUF, S. FISCHER, H. BUNKE, M.J.F. WILKINSON, J.B.T.M. ROERDINK, J. PECH-PACHECO, G. CRISTOBAL, V. CIRIMELE & B. LUDES, B. (2001): ADIAC: using computer vision technology for automatic diatom identification. – In: ECONOMOU-AMILLI, A. (ed.): Proceedings of the 16th International Diatom Symposium: 537–562. Faculty of Biology, University of Athens, Greece.

BEAUFORT, L. & D. DOLLFUS (2004): Automatic recognition of coccoliths by dynamical neural networks. – Mar. Micropaleontol. **51**: 57–73.

BEHNKE, A., T. FRIEDL, V.A. CHEPURNOV & D.G. MANN (2004): Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). – J. Phycol. **40**: 193–208.

BLAXTER, M.L. (2004): The promise of a DNA taxonomy. – Phil. Trans. R. Soc. Lond., **B 359**: 669–679.

BLAXTER M., B. ELSWORTH & J. DAUB (2004): DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. Proc. R. Soc. Lond., **B 271**: S189–S192.

BOLLMANN, J., P.S. QUINN, M. VELA, B. BRABEC, S. BRECHNER, M.Y. CORTÉS, H. HILBRECHT, D.N. SCHMIDT, R. SCHIEBEL & H.R. THIERSTEIN (2004, in press). Automated particle analysis: calcareous microfossils. – In: FRANCUS, P. (ed.): Image Analysis, Sediments and Paleoenvironments (Developments in Paleoenvironmental Research, vol. 7). Kluwer, Dordrecht.

BOOKSTEIN, F.L. (1991): Morphometric tools for landmark data: geometry and biology. – Cambridge University Press, Cambridge.

CAIRNS, J. JR (2002): Environmental monitoring for the preservation of global biodiversity the role in sustainable use of the planet. – Int. J. Sustain Dev. World Ecol. **9**: 135–150.

CAIRNS, J. JR, R.L. KAESLER & R. PATRICK (1970): Occurrence and distribution of diatoms and other algae in the upper Potomac River. – Notulae naturae **436**: 1–12.

CAIRNS, J. JR, K.L. DICKSON, G.R. LANZA, S.P. ALMEIDA & D. DEL BALZO (1972): Coherent optical spatial filtering of diatoms in water pollution monitoring. – Arch. Mikrobiol. **83**: 141–146.

CAIRNS, J. JR, K.L. DICKSON & G.R. LANZA (1973): Rapid biological monitoring systems for determining aquatic community structure in receiving systems. – In: CAIRNS, J. JR, & K.L. DICKSON (eds): Biological methods in the assessment of water quality: 148–163. American Society for Testing and Materials, Philadelphia, STP 528.

CAIRNS, J. JR, J.W. HALL, E.L. MORGAN, R.E. SPARKS, W.T. WALLER & G.F. WESTLAKE (1974a): The development of an automated biological monitoring system for water quality management. – In: Proceedings of the 7th Annual Conference on Trace Substances in Environmental Health: 43–55. University of Missouri, Columbia, Missouri.

CAIRNS, J. JR, K.L. DICKSON, J.P. SLOCOMB, S.P. ALMEIDA, J.K.T. EU, C.Y.C. LIU & H.F. SMITH (1974b): Microcosm pollution monitoring. – In: Hemphill, D.D. (ed.) Proceedings of the 8th Annual Conference on Trace Substances in Environmental Health: 233–228. University of Missouri, Columbia, Missouri.

CAIRNS, J. JR, K.L. DICKSON & J. SLOCOMB (1977a): The ABC's of diatom identification using laser holography. – Hydrobiologia **54**: 7–16.

CAIRNS, J. JR, K.L. DICKSON & J. SLOCOMB, S.P. ALMEIDA & J.K.T. EU (1977b): Biological monitoring of aquatic community structure using a computer interfaced laser system. – In: ALABASTER, J.S. (ed.): Biological monitoring of inland fisheries: 143–150. Applied Science Publishers, London.

CAIRNS, J. JR, K.L. DICKSON, P. PRYFOGLE, S.P. ALMEIDA, S.K. CASE, J.M. FOURNIER & H. FUJII (1979): Determining the accuracy of coherent optical identification of diatoms. – Water Resource Bull. **15**: 1770–1775.

CAIRNS, J. JR, S.P. ALMEIDA & H. FUJII (1982): Automated identification of diatoms. – BioScience **32**: 98–102.

CASE, S.K., S.P. ALMEIDA, W.J. DALLAS, J.M. FOURNIER, K. PRITZ, J. CAIRNS JR, K.L. DICKSON & P.A. PRYFOGLE (1978): Coherent microscopy and matched spatial filtering for real-time recognition of diatom species. – Environm. Sci. Technol. **12**: 940–946.

CIOBANU, A. & H. DU BUF (2002): Identification by contour profiling and Legendre polynomials. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 167–185. – World Scientific Publishing Co., Singapore.

CIOBANU, A., H. SHAHBAZKIA & J.M.H. DU BUF (2000): Contour profiling by dynamic ellipse fitting. – In: SANFELIU A., J.J. VILLANUEVA, M. VANRELL, R. ALQUEZAR, T. HUANG & J. SERRA (eds): Proceedings, 15th International Conference on Pattern Recognition (ICPR 2000): 750-753. IEEE Computer Society.

DAVIDOVICH, N.A. & S.S. BATES (1998): Sexual reproduction in the pennate diatoms *Pseudo-nitzschia multiseries* and *P. pseudodelicatissima* (Bacillariophyceae). – J. Phycol. **34**, 126–137.

DICKSON, K.L., J.P. SLOCOMB, J. CAIRNS JR, S.P. ALMEIDA & J.K.T. EU (1976): A laser based optical filtering system to analyze samples of diatom communities. – In: CAIRNS, J. JR, K.L. DICKSON & G.F. WESTLAKE (eds): Biological monitoring of water and effluent quality. American Society for Testing and Materials, Philadelphia. Special Technical Publication 607.

DROOP, S.J.M. (1994): Morphological variation in *Diploneis smithii* and *D. fusca* (Bacillariophyceae). – Arch. Protistenk. **144:** 249–270.

DROOP, S.J.M., D.G. MANN & G.M. LOKHORST (2000): Spatial and temporal stability of demes in *Diploneis smithii/D. fusca* (Bacillariophyta) supports a narrow species concept. – Phycologia **39**: 527–546.

DU BUF, J.M.H. & M.M. BAYER (eds.) (2002a): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51. – World Scientific Publishing Co., Singapore.

DU BUF, J.M.H. & M.M. BAYER (2002b): Introduction to ADIAC and this book. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 1–8. – World Scientific Publishing Co., Singapore.

DU BUF, H., M. BAYER, S. DROOP, R. HEAD, S. JUGGINS, S. FISCHER, H. BUNKE, M. WILKINSON, J. ROERDINK, J. PECH-PACHECO & G. CRISTOBAL (2000). Diatom identification: a double challenge called ADIAC. – In: Proceedings of the 10th International Conference on Image Analysis and Processing, Venice, Italy, September 27–29, 1999: 734–739.

ELORANTA, P. (1987): Fritsch Algae Collection. Genus guide to the Basic Collection and to Supplements I–IV. – Inter Documentation Company, Leiden, The Netherlands.

FISCHER, S. & H. BUNKE (2001): Automatic identification of diatoms using decision forests. – Lecture Notes Artificial Intelligence. **2123**: 173–183.

FISCHER, S. & H. BUNKE (2002a): Automatic identification of diatoms using visual human-interpretable features. – Int. J. Image and Graphics 2: 67–87.

FISCHER, S. & H. BUNKE (2002b): Identification using classical and new features in combination with decision tree ensembles. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 109–140. – World Scientific Publishing Co., Singapore.

FISCHER, S., M. BINKERT & H. BUNKE (2000a): Symmetry based indexing of diatoms in an image database. – In: Proceedings, 15th International Conference on Pattern Recognition (ICPR 2000): 2895–2898. IEEE Computer Society.

FISCHER, S., M. BINKERT & H. BUNKE (2000b): Feature based retrieval of diatoms in an image database using decision trees. – In: Proceedings of the 2nd International Symposium on Advanced Concepts for Intelligent Vision Systems (ACIVS 2000): 67–72.

FISCHER, S., K. GILOMEN & H. BUNKE (2002): Identification of diatoms by grid graph matching. – Lecture Notes Computer Sci. 2396: 94–103.

FORERO-VARGAS, M., R. REDONDO & G. CRISTOBAL (2003): Diatom screening and classification by shape analysis. – Lecture Notes Computer Sci. **2849**: 58–65.

FRANCE, I., A.W.G. DULLER, H.F. LAMB & G.A.T. DULLER (1997): A comparative study of approaches to automatic pollen identification. – Proc. British Machine Vision Conference **1997**: 340–349.

FRANCE, I., A.W.G. DULLER, G.A.T. DULLER & H.F. LAMB (2000): A new approach to automated pollen analysis. – Quatern. Sci. Rev. **19**: 537–546.

FRANCE, I., A.W.G. DULLER & G.A.T. DULLER (2004, in press): Software aspects of automated recognition of particles: the example of pollen. – In: FRANCUS, P. (ed.): Image analysis, sediments and paleoenvironments (Developments in Paleoenvironmental Research, vol. 7). Kluwer, Dordrecht.

FUJII, H. & S.P. ALMEIDA (1979): Coherent spatial filtering with simulated input. – Appl. Optics **18**: 1659–1662.

FUJII, H., S.P. ALMEIDA & J.E. DOWLING (1980): Rotational matched spatial filter for biological pattern recognition. – Appl. Optics **19**: 1190–1195.

FUKUSHIMA, H., S. YOSHITAKE & T. KO-BAYASHI (2001): *Pinnularia paralange-bertalotii* Fukush., Yoshit. & Ts. Kobay. nov. spec., a new diatom from acid water. – Diatom **17**: 37–46.

FUKUSHIMA, H., S. YOSHITAKE & T. KO-BAYASHI (2004): Morphological variability of *Pinnularia osoresanensis* (Negoro) Fukush., Yoshit. & Ts. Kobay. from Osorezan, northern Japan. – In: Poulin, M. (ed.): Proceedings of the 17th International Diatom Symposium: 93–102. Biopress, Bristol.

GASTON, K.J. & M.A. O'NEILL (2004): Automated species identification: why not? – Phil. Trans. R. Soc. Lond., **B 359**: 655–667.

HEBERT, P.D.N., A. CYWINSKA, S.L. BALL & J.R. DEWAARD (2003a): Biological identifications through DNA barcodes. – Proc. R. Soc. Lond., **B 270**: 313–321.

HEBERT, P.D.N., S. RATNASINGHAM, & J.R. DEWAARD (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. – Proc. R. Soc. Lond., **B 270**: S96-S99.

HICKS Y., D. MARSHALL, R.R. MARTIN, P.L. ROSIN, M.M. BAYER & D.G. MANN (2002). Automatic landmarking for building biological shape models. – In: Proceedings of the 2002 IEEE International Conference on Image Processing (ICIP 2002), vol 2: 801–804. IEEE Computer Society.

HICKS, Y., D. MARSHALL, R.R. MARTIN, P.L. ROSIN, S. DROOP & D.G. MANN (2004, in press): Building shape and texture models of diatoms for analysis and synthesis of drawings and identification. – In: Proceedings of the Irish Machine Vision and Image Processing Conference 2004 (IMVIP 2004)

HUSTEDT, F. (1937): Zur Systematik der Diatomeen. II. Der Begriff des "Typus" bei den Diatomeen und der Umfang der Diagnosen. III. Konvergenzerscheinungen und Kümmerformen. – Ber. deutsch. bot. Ges. **55:** 465–472.

JAHN, R., W.-H. KUSBER, L.K. MEDLIN, R.M. CRAWFORD, D. LAZARUS, T. FRIEDL, D. HEPPERLE, B. BESZTERI, K. HAMANN, F. HINZ, S. STRIEBEN, V. HUCK, J. KASTEN, A. JOBST & K. GLÜCK (2004). Taxonomic, molecular and ecological information on diatoms: the information system *AlgaTerra*. – In: Poulin, M. (ed.): Proceedings of the 17th International Diatom Symposium: 121–128. Biopress, Bristol.

JALBA, A.C. & J.B.T.M. ROERDINK (2003): Automatic segmentation of diatom images. – Lecture Notes Computer Sci. **2756**: 329–336.

JALBA, A.C., M.H.F. WILKINSON & J.B.T.M. ROERDINK (2004): Morphological hat-transform scale spaces and their use in pattern classification. – Pattern Recognition **37**: 901–915.

KELLY, M.G., M.M. BAYER, J. HÜRLIMANN & R.J. TELFORD (2002): Human error and quality assurance in diatom analysis. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 75–91. – World Scientific Publishing Co., Singapore

KRAMMER, K. (1992): *Pinnularia*: eine Monographie der europäischen Taxa. – Bibliotheca diatomol. **26**: 1–353.

KRAMMER, K. (2000): The genus *Pinnularia*. In: LANGE-BERTALOT, H. (ed.): Diatoms of Europe, vol. 1. A.R.G. Gantner, Ruggell, Liechtenstein.

KRAMMER, K. & H. LANGE-BERTALOT (1986): Bacillariophyceae 1. Teil: Naviculaceae. – In: ETTL, H., J. GERLOFF, H. HEYNIG & D. MOLLENHAUER (eds): Süsswasserflora von Mitteleuropa, vol. 2/1. G. Fischer, Stuttgart & New York.

KRAMMER, K. & H. LANGE-BERTALOT (1988): Bacillariophyceae 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. – In: ETTL, H., J. GERLOFF, H. HEYNIG & D. MOLLENHAUER (eds): Süsswasserflora von Mitteleuropa, vol. 2/2. G. Fischer, Stuttgart & New York

KRAMMER, K. & H. LANGE-BERTALOT (1991a): Bacillariophyceae 3. Teil: Centrales, Fragilariaceae, Eunotiaceae. – In: Ettl, H., J. GERLOFF, H. HEYNIG & D. MOLLENHAUER (eds): Süsswasserflora von Mitteleuropa, vol. 2/3. G. Fischer, Stuttgart & New York.

KRAMMER, K. & H. LANGE-BERTALOT (1991b): Bacillariophyceae 4. Teil: Achnanthaceae. Kritische Ergänzungen zu *Navicula* (Lineolatae) und *Gomphonema*. – In: ETTL, H., G. GÄRTNER, J. GERLOFF, H. HEYNIG & D. MOLLENHAUER (eds): Süsswasserflora von Mitteleuropa, vol. 2/4. G. Fischer, Stuttgart & New York.

KRÖGER, N., R. DEUTZMANN & M. SUMPER (1999): Polycationic peptides from diatom biosilica that direct silica nanosphere formation. – Science **286**: 1129–1132.

KRÖGER, N., S. LORENZ, E. BRUNNER & M. SUMPER (2002): Self-assembly of highly phosphorylated silaffins and their function in biosilica morphogenesis. – Science **298**: 584–586.

LANGE-BERTALOT, H. & D. METZELTIN (1996): Oligotrophie-Indikatoren. 800 Taxa repräsentative für drei diverse Seen-Typen; kalkreich–oligodystroph–schwach gepuffertes Weichwasser. – Iconographia diatomol. **2:** 1-390.

LAZARUS, D. (1998): The Ehrenberg Collection and its curation. – *Linnean, Special Issue* **1:** 31–48.

LOKE, R.E. & H DU BUF (2002): Identification by curvature of convex and concave segments. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 141–165. – World Scientific Publishing Co., Singapore.

LOKE, R.E., M.M. BAYER, D.G. MANN & J.M.H. DU BUF (2002): Diatom recognition by convex and concave contour curvature. – In: Proceedings of the Oceans '02 MTS/IEEE Conference: 2457-2465. IEEE Computer Society.

LOKE, R.E., J.M.H. DU BUF, M.M. BAYER & D.G. MANN (2004): Diatom classification in ecological applications. – Pattern Recognition 37: 1283–1285.

LUNDHOLM, N., Ø. MOESTRUP, G.R. HASLE & K. HOEF-EMDEN (2003): A study of the *Pseudo-nitzschia pseudodelicatissima/cuspidata* complex (Bacillariophyceae): what is *P. pseudodelicatissima*? – J. Phycol. **39**: 797–813.

MANN, D.G. (1988): Why didn't Lund see sex in *Asterionella*? A discussion of the diatom life cycle in nature. – In ROUND, F.E. (ed.): Algae and the Aquatic Environment: 383–412. – Biopress, Bristol.

MANN, D.G. (1998): Ehrenbergiana: problems of elusive types and old collections, with special reference to diatoms. – *Linnean, Special Issue* **1:** 63–88.

MANN, D.G. (1999): The species concept in diatoms (Phycological Reviews 18). – Phycologia **38**: 437–495.

MANN, D.G. & S.J.M. DROOP (1996): Biodiversity, biogeography and conservation of diatoms. – Hydrobiologia **336**: 19–32.

MANN, D.G., S.M. MCDONALD, M.M. BAYER, S.J.M. DROOP, V.A. CHEPURNOV, R.E. LOKE, A. CIOBANU, & J.M.H. DU BUF (2004): Morphometric analysis, ultrastructure and mating data provide evidence for five new species of *Sellaphora* (Bacillariophyceae). – Phycologia **43**: 459-482.

MEDLIN, L.K. & I. KACZMARSKA (2004): Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. – Phycologia **43**: 245–270.

MOU, D. & E.F. STOERMER (1992): Separating *Tabellaria* (Bacillariophyceae) shape groups based on Fourier descriptors.– J. Phycol. **28:** 386–395.

NIPKOW, F. (1927): Über das Verhalten der Skelette planktischer Kieselalgen im geschichteten Tiefenschlamm des Zürich- und Baldeggersees. – *Z. Hydrogr. Hydrobiol.* **4:** 71–120.

PAPPAS, J.L. & E.F. STOERMER (2003): Legendre shape descriptors and shape group determination of specimens in the *Cymbella cistula* species complex. – Phycologia **42**: 90–97.

PAPPAS, J.L., G.W. FOWLER & E.F. STOERMER (2001): Calculating shape descriptors from Fourier analysis: shape analysis of *Asterionella* (Heterokontophyta, Bacillariophyceae). Phycologia **40**: 440–456.

PARTIN, J.K., S.P. ALMEIDA & H. FUJII (1979): Use of an interference contrast microscope in a coherent optical processor. – Proc. Soc. Photo-Optical Instrum. Eng. Conf. **177**: 50–53.

PATRICK, R. (1967): The effect of invasion rate, species pool, and size of area on the structure of the diatom community. – Proc. Natl. Acad. Sci. USA **58**: 1335–1342.

PATRICK, R., M.H. HOHN & J.H. WALLACE (1954): A new method for determining the pattern of the diatom flora. – Notulae naturae **259**: 1–12.

PECH-PACHECO, J. & G. CRISTÓBAL (2002): Automatic slide scanning. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 259–288. – World Scientific Publishing Co., Singapore

PICKETT-HEAPS, J.D., A.-M.M. SCHMID & L.A. EDGAR (1990): The cell biology of diatom valve formation. – Progr. Phycol. Res. **7**: 1–168.

RHODE, K.M., J.L. PAPPAS & E.F. STOERMER (2001): Quantitative analysis of shape variation in type and modern populations of *Meridion* (Bacillariophyceae). – J. Phycol. **37**: 175–183.

ROSIN, P.L. (2003): Measuring shape: ellipticity, rectangularity, and triangularity. – Machine Vision Applic. 14: 172–184.

ROSIN, P.L. (2004): Measuring sigmoidality. – Pattern Recognition **37**: 1735–1744.

ROUND, F.E., R.M. CRAWFORD & D.G. MANN (1990): The diatoms. Biology and morphology of the genera. – Cambridge University Press, Cambridge.

SANTOS, L. & H. DU BUF (2002): Identification by Gabor features. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 187–220. – World Scientific Publishing Co., Singapore.

SCHMIDT, A., M. SCHMIDT, F. FRICKE, H. HEIDEN, O. MÜLLER & F. HUSTEDT (1874–1959): Atlas der Diatomaceen-Kunde. – O.R. Reisland, Leipzig.

SCOTLAND, R., C. HUGHES, D. BAILEY & A. WORTLEY (2003): The *Big Machine* and the much-maligned taxonomist. – Syst. Biodivers. **1**: 139–143.

STOERMER, E.F. & T.B. LADEWSKI (1982): Quantitative analysis of shape variation in type and modern populations of *Gomphoneis herculeana*. – Nova Hedwigia, Beiheft **73:** 347-386.

STOERMER, E.F. & J.P. SMOL (eds) (1999): The diatoms: applications for the environmental and earth sciences. – Cambridge University Press, Cambridge.

WESTENBERG, M.A. & J.B.T.M. ROERDINK (2002): Mixed-method identifications. – – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 245–257. – World Scientific Publishing Co., Singapore.

WILKINSON, M.H.F., J.B.T.M. ROERDINK, S. DROOP & M. BAYER (2000): Diatom contour analysis using morphological curvature scale spaces. In: SANFELIU A., J.J. VILLANUEVA, M. VANRELL, R. ALQUEZAR, T. HUANG & J. SERRA (eds): Proceedings of the 15th International Conference on Pattern Recognition, vol 3: 652–655.

WILKINSON, M.H.F., A.C. JALBA, E.R. URBACH & J.B.T.M. ROERDINK (2002). Identification by mathematical morphology. – In: DU BUF, J.M.H. & M.M. BAYER (eds): Automatic diatom identification. Series in Machine Perception and Artificial Intelligence, vol. 51: 221–244. – World Scientific Publishing Co., Singapore.

WILL, K.W. & D. RUBINOFF (2004): Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. – Cladistics **20**: 47–55.

YU S., P. SAINTMARC, M. THONNAT & M. BERTHOD (1996): Feasibility study of automatic identification of planktic foraminifera by computer vision. – J. Foramin. Res. **26**: 113–123.

ZUNIC, J. & P.L. ROSIN (2004): A new convexity measure for polygons. – IEEE Trans. Pattern Anal. Machine Intell. 26: 923–934.

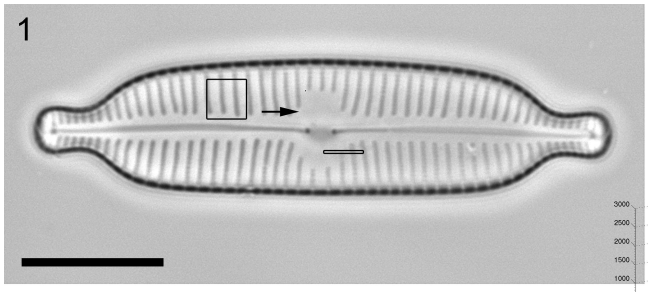
Table 1. Feature sets tested in the ADIAC project

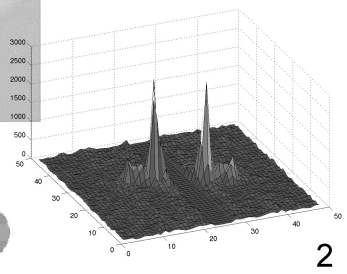
Type of feature	Features
Symmetry	class of symmetry (heteropolar, dorsiventral, circular, etc); strength of apical and transapical symmetry
Overall shape descriptors	rectangularity, circularity, ellipticity, triangularity, compactness
Shape properties	global shape properties (via polynomial approximation of the centre line)
	Pole shape
	Fourier descriptors
	Moment invariants
	Contour segment analysis
	Contour profiling
	Gabor contour features
	Legendre polynomials
	Multi-scale morphological analysis by 1–D hat transform
Geometrical properties and size	length, width, length–width ratio, area inside contour
Rib and stria pattern	stria density, orientation, changeover point
	axial area width
	costa density (e.g. *Denticula*, *Diatoma*)
	frequency of structural elements longitudinally (where striae do not form continuous lines)
Raphe	presence (use of Gabor bar cell)
Texture	grey-level co-occurrence (characterizes the frequencies with which particular grey levels occur in particular spatial relationships to other grey levels)
	Gabor wavelets (use of a bank of Gabor filters, ‘sensitive’ to particular scales, orientations and types of pattern)
	Multi-scale morphological analysis by 2–D hat transform

Figure legends

Figs 1–7. *Cymbella hybrida*. Fig. 1. Digital micrograph, showing the sizes of window used for characterizing the stria pattern (square window) and for detecting the axial area (elongate window). The window is centred on each pixel in turn and pattern characterized by Fast Fourier Transform. Scale bar = 10 μm. Fig. 2. Example of 2-D FFT output. The positions of the peaks indicate the orientation and spacing of repeated pattern, and the heights (energy) are a measure of pattern intensity. Fig. 3. Square window FFTs: map of stria frequency (higher stria densities are darker) for the valve shown in Fig. 1. Fig. 4. Square window FFTs: map of stria orientation (more positive gradients are darker). Fig. 5. Elongate window FFTs: energy map (low energy is dark and shows relative absence of pattern). Fig. 6. Axial and central areas detected by thresholding the map in Fig. 5. Fig. 7. Drawing of the valve shown in Fig. 1, synthesized from a 301-element vector describing shape, stria pattern, axial area shape, and size.

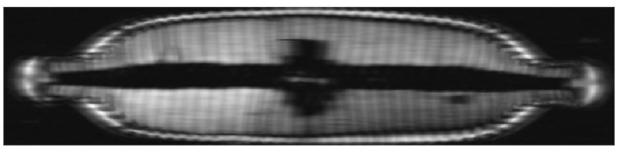
Figs 8–10. *Gomphonema* cf. *minutum*. Fig. 8. Four of the input images used to develop a model of variation in *G.* cf. *minutum*. Scale bar = 10 μm. Fig. 9. The fitted Principal Curve for *G.* cf. *minutum*, based on the 301-element shape–pattern vectors for 19 valves, projected into the space of the first three Principal Components. Fig. 10. Drawings of three ‘virtual specimens’ of *G.* cf. *minutum*, derived from nodes along the Principal Curve. Note that these do not correspond to any of the actual specimens used to develop the model but are constructed from the model itself. Note also that the model accurately reflects a slight lateral asymmetry in the valves, which are slightly flatter (less convex) on the secondary side (opposite the stigma: on the left in the largest valve in Fig. 8 and on the right in the others).

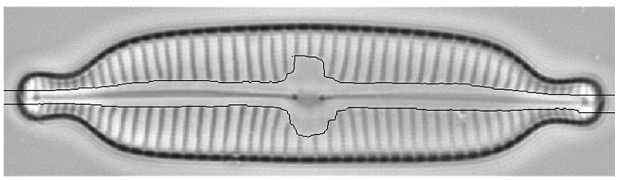

1

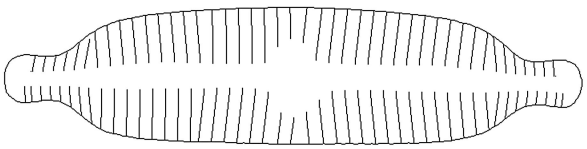

3000
2500
2000
1500
1000
500
0
50
40
30
20
10
0
0
10
20
30
40
50
2

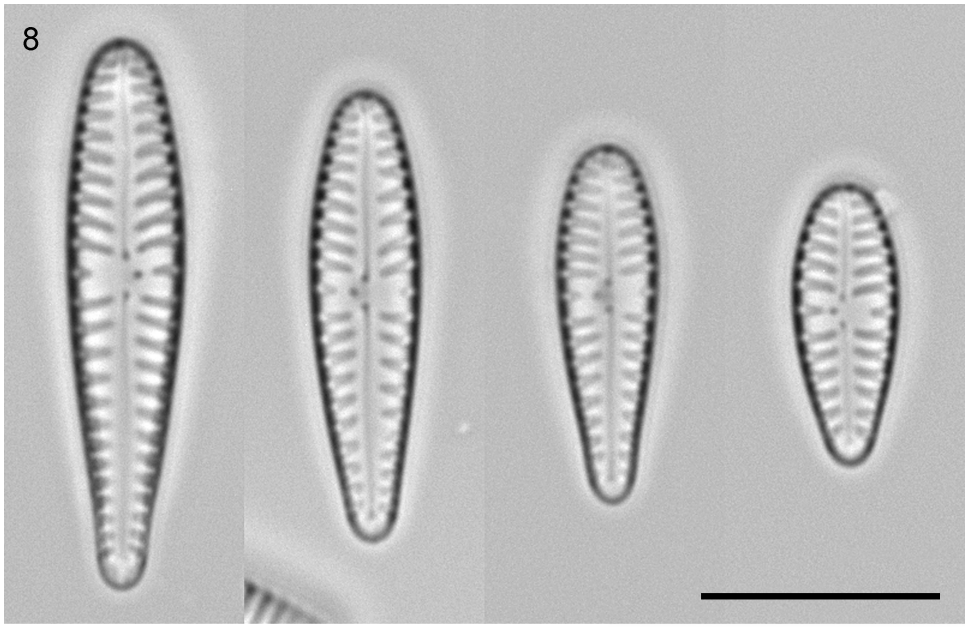

3

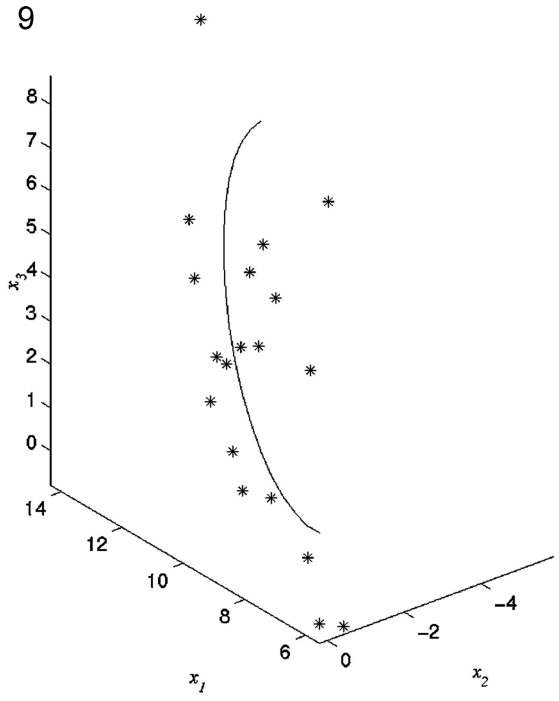

4

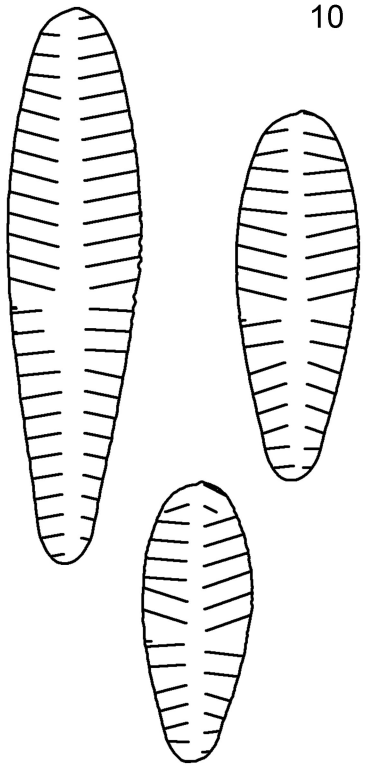

5


6


7


8


9
8
7
6
5
4
3
2
1
0
x_3
14
12
10
8
6
x_1
0
−2
−4
x_2


10