Fourier transforms of image-derived data: application to Albian coccoliths

Jane Garratt and Andrew R. H. Swan
School of Geological Sciences, Kingston University, Penrhyn Road, Kingston-upon-Thames, Surrey, KT1 2EE, UK

ABSTRACT: The reduction of image-derived morphological data to a suite of characters that can be regarded as homologous is problematic. On the basis of analyses of hypothetical artificial morphologies, it is argued that the principal components of Fourier transforms of radially symmetrical structures are correlated with morphological attributes identified *a priori* as homologous. Application to grey-level images of Albian coccoliths yields results demonstrating that successful discrimination of genera and species is possible using an automatic retrieval and analysis system.

A REVIEW OF FOURIER ANALYSIS IN PALEONTOLOGY

Over the last ten years interest has grown in the use of image analysis techniques to retrieve information on fossil shapes. This information can, in principle, be used to quantify evolutionary changes within taxa and to introduce objectivity into taxonomy, biostratigraphy and the identification of specimens. A fundamental objective is the reduction of the large quantity of information inherent in an image to a small set of relevant morphological descriptors. Fourier analysis, mainly applied to outlines, is one approach to this problem: it can be argued that Fourier harmonics are efficient shape descriptors and that the power spectrum allows data reduction.

Modern morphological studies using these techniques are based on three early studies. Benson (1967) showed that morphological data could be expressed as polar coordinates and analyzed using the "theta-rho" technique; Schwartz and Shane (1969) described how closed boundaries could be represented by Fourier harmonics; and Ehrlich and Weinberg (1970) extended this technique to obtain precise estimates, and reproductions, of shapes using Fourier harmonic amplitudes.

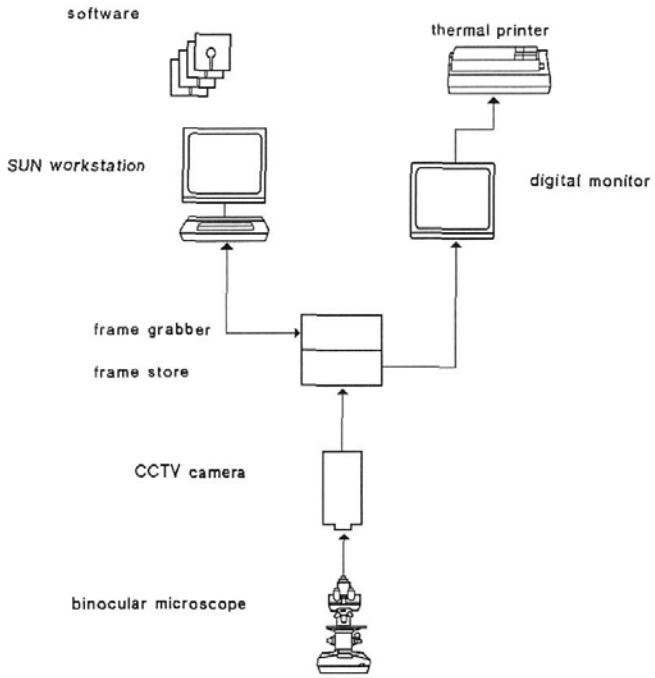
Subsequent research has investigated diverse taxa including bryozoans (Anstey and Delmet 1973; Delmet and Anstey 1974); ostracodes (Kaesler and Waters 1972; Burke *et al.* 1987); foraminifera (Healy-Williams and Williams 1981; Healy-Williams 1983, 1984; Lohmann 1983; Malmgren *et al.* 1984); blastoids (Waters 1977); miospores (Christopher and Waters 1974) and trilobites (Foote 1989, 1991). These studies all use the outline shape of the specimen as a basis for analysis. The Fourier technique was extended to the analysis of the internal morphologies of coccolith images (Garratt and Swan 1993).

THE PROBLEM OF HOMOLOGY

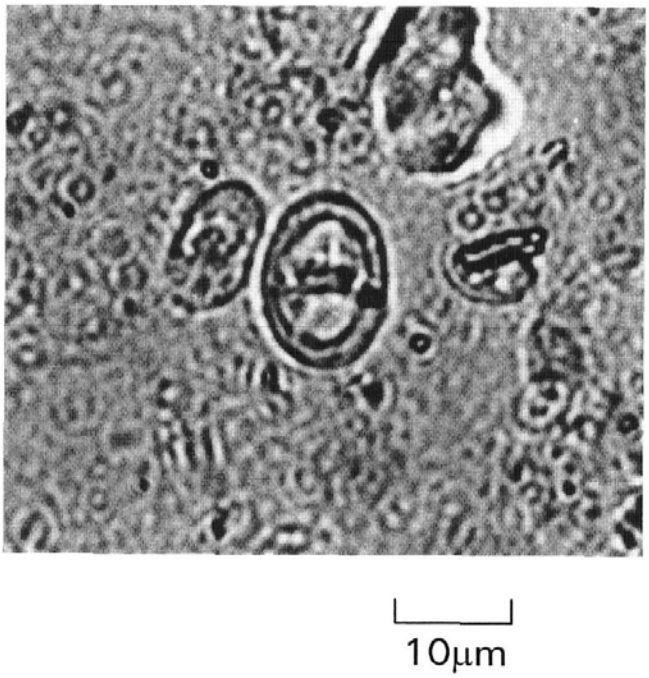
In morphometric analyses based on images, the successful interpretation of the results depends on the derivation of homologous and biologically meaningful characters from large arrays of image pixel values which are not, apparently, either homologous or biologically meaningful.

An accepted general definition of homology is "the possession by two or more species of a trait derived, with or without modification, from their common ancestor" (Wagner 1989). Two difficulties arise from this: firstly, there is a problematic requirement of knowledge of phylogeny; secondly, the idea of degrees of modification allows more than one interpretation. "Taxic homology" (Eldredge 1979) is equivalent to synapomorphy, and cladistic procedure regards synapomorphic character states in two organisms as identical. In "operational homology" (the "quantitative" definition of Sneath and Sokal 1973), homology is identified on the basis of structural correspondence and characters can be identified as homologous if they are found by applying the same set of structural criteria to each specimen. Although this can proceed with no knowledge of phylogeny, the intention is to find attributes that are homologous in the phylogenetic sense. In practice, operational homology is often invoked in morphometrics because structurally corresponding characters may be quantitatively different and comparison of numerical values can be justified.

As operational homology reduces the criterion of descent to one of interpretation, the opportunity arises for homology to be claimed on feeble evidence. Indeed, there is a continuum of validity between phylogenetically justified homology and *ad hoc* replicable measurement. Collection of morphometric data is often undertaken on the basis that the measurements on different specimens are done using the same operational procedure: measurements of the same feature or dimension are entered in the same column in the data set and all values of this variable are manipulated in the same way and all are regarded as mathematically comparable (e.g. differences obtained by subtraction are meaningful). Clearly, however, homology is a property which is separate from mathematical comparability of data. Full and Ehrlich (1982) observed that, since two identical shapes would produce identical Fourier series, the amplitude spectra for two non-identical shapes can be validly compared because "each term of equivalent order represents the 'same thing' in each series". These terms are mathematically comparable since the measurements are made in the same manner and record the same type of data, but they are not necessarily homologous. Full and Ehrlich (1986), in their text-figure 2, demonstrate this problem: the authors correctly state that mathematically comparable



TEXT-FIGURE 1
Schematic diagram of the image analysis system used in this study.



TEXT-FIGURE 2
A typical digitised image, reproduced by the thermal printer.

measurements made at the same physical points on two outlines may not be homologous. Ferson et al. (1985) used a Fourier method to analyse *Mytilus* outlines and acknowledged that it is "fruitless to attach biological interpretations to Fourier coefficients" but affirmed the usefulness of shape analysis even when isolated from considerations of homology.

Bookstein (1994) argued on theoretical grounds that "Morphometrics cannot supply homologous shape characters but must be informed about them in advance": this appears to concede that a morphometric measurement procedure can be designed to strive to retrieve homologous data but that data processing cannot enhance homology. However, he also asserts that, although some microevolutionary trends may be amenable to morphometric analysis, "no important evolutionary change can be captured persuasively in the language of biometrics". There is little doubt from Bookstein's mathematical considerations that we cannot expect a linear relationship between morphometric parameter values and truly homologous genetic information.

The position adopted in the present study with regard to Bookstein's (1994) argument is as follows:

1. Taxic homology can very rarely be demonstrated and certainly not with the type of material (Albian coccoliths) considered here; consequently, we will attempt sensible *a priori* conjectures as to which characters may be regarded as homologous. Such conjecture is open to debate and revision, in the same way as any other paleontological hypothesis. Morphometric data will then be collected and processed in order to quantify the selected characters.

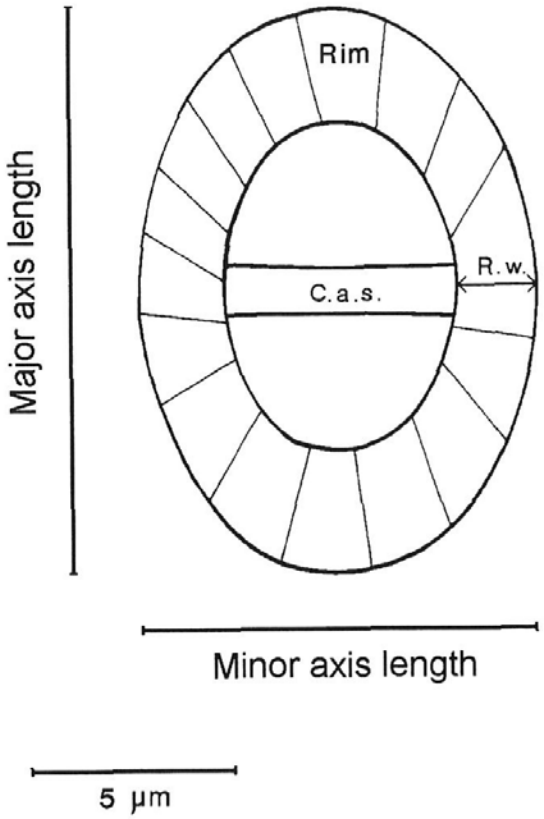
2. Bookstein's discussion does not acknowledge the shades of quality and uncertainty in taxonomic methods and results. There are a variety of ways of quantifying the shape of any biological object (e.g. image pixel values, landmarks, simple dimensions, Fourier transforms): these vary in efficiency of description and clearly have varying correlations with genotypic information. It is always reasonable to collect and process data such that these properties can be claimed to be enhanced. Furthermore, a morphometric result is no different from any other scientific result in being subject to revision or refutal. Even if confirmed, a morphometric result only expresses quantitative relationships between data, but this may lead, via interpretation, to hypotheses about taxonomic relationships or evolutionary trends. These in turn may prove to be true or false, but it should be expected that the result will be a useful contribution to the body of knowledge of the group in question.

DATA FROM COCCOLITH IMAGES

In this study, the problems of homology are to be investigated in the context of data derived from Albian (Cretaceous) coccolith images. These were retrieved from smear slides using an oil immersion lens and a light microscope to produce a magnification factor of x1000. The images were sent to a SUN workstation and were held in a 768 x 576 pixel raster array for display on a monitor screen. Typical coccoliths have long axis lengths of between 60 and 200 pixels depending on genus. They were then processed by storing and analysing the grey-level data that comprise the images. Text-figure 1 shows the image analysis system used, and text-figure 2 shows a typical coccolith image.

Homologous characters were identified in advance of the analysis on the basis of their general acceptance in existing taxonomy. These characters are of three types:-

A. Major dimensions.
1). Length of the major axis of the coccolith ellipse.
2). Ratio of the major to the minor axes.



TEXT-FIGURE 3
A simplified sketch of a typical specimen of the genus *Zeugrhabdotus* showing the coccolith elements used in the analysis. Legend: C.a.s = Central area structure. R.w. = Rim width.

3). Ratio of the rim width to the total width of the specimen.

B. Rim structures

4). Radial symmetry of the coccolith rim.

C. Central area structures.

5). Radial symmetry of the central area structure.

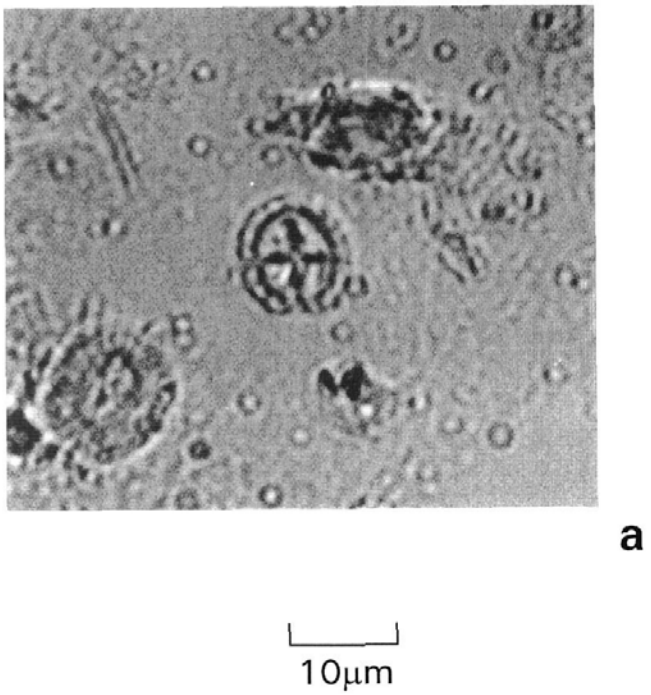
Characters 1 and 2 are taken from direct measurements of the image, whilst characters 3, 4 and 5 are derived from the grey-level data. Characters 4 and 5 are represented by multiple variables in the analysis. The condensation of the information on radial symmetry is the main issue in the analytical procedures. Text-figure 3 illustrates these characters.

Coccolith specimens from the genera *Prediscosphaera* and *Zeugrhabdotus* were used in this case study (text-figure 4). It can be seen that specimens of the genus *Prediscosphaera* are roughly circular, contain a central cross and have about 16 rim plates. Specimens of the genus *Zeugrhabdotus* tend to be larger and elliptical in shape, have a single bar in the central area and many ill-defined rim plates.

BACKGROUND TO THE ANALYTICAL TECHNIQUES

Data retrieval

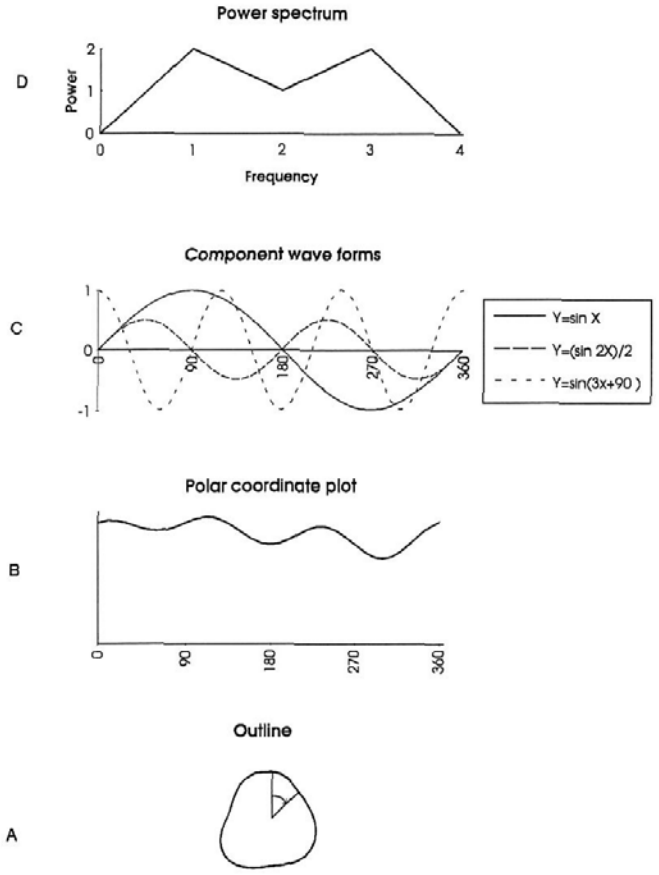
Greylevel data were retrieved using concentric elliptical scans of the pixel array, with 128 positions at regular angular increments on each scan, and with the radius increased in 40 increments from near 0 to the coccolith outline. The scans were centred manually. The grey-level value of the pixel in each position in the scans is stored and used as input to the next stage of processing.



TEXT-FIGURE 4
Images of specimens of *Prediscosphaera columnata* (Text-figure 4a) and *Zeughrhabdotus* (Text-figure 4b) as produced by the system.

Fourier analysis

Fourier analysis breaks down the variation within a time series into component parts according to the duration of the intervals within which the variation occurs, and is valid if the time series is periodic. The total variance of the Fourier time series is composed of the sum of the variances of its component harmonics where each harmonic is expressed as a cosine wave with the specified frequency, amplitude and phase.



TEXT-FIGURE 5
The different stages in the production of a power spectrum. The object outline is scanned (Text-figure 5A), producing a series of measurements which can be unrolled as a polar coordinate plot (Text-figure 5B). This wave form is resolved into its component harmonics by the Fourier transform (Text-figure 5C), and the power spectrum for these harmonics is calculated and displayed graphically (Text-figure 5D).

In this analysis the input data are derived from the elliptical scans of the grey-levels in coccolith images. These are analogous to the polar coordinates used in outline analysis and the same considerations are applicable to both; consequently, the circular form of the Fourier transform, which uses polar coordinates, was used (Rohlf 1990).

Output from the Fourier Transform can be numeric or displayed in the form of a periodogram or power spectrum. This shows the proportion of the total variance in the original time series which is contributed by each component wave-form. Text-figure 5 shows the stages of production of a power spectrum using this procedure. An object is scanned (text-figure 5a), retrieving data that can be "unwrapped" to construct a wave form (text-figure 5b). This is then resolved into its component harmonics (text-figure 5c) and the power spectrum for these harmonics is calculated and displayed graphically (text-figure 5d).

To resolve the data 40 power spectra are produced for each image, representing the internal coccolith structures found at increasing radii from the coccolith centre. For ease of interpretation, all the power spectra produced for one image are displayed on a single, composite, diagram.

Text-figure 6 shows the data retrieved from a specimen of *Prediscosphaera columnata*. On these composite diagrams the centre of the coccolith is at the base. For each scan, the vertical axis is power, but the scans are separated according to distance from the centre for ease of interpretation. Within the power spectrum, for each separate radius, the peaks indicate frequencies at which there are cyclic variations in grey-level. The frequencies at which these peaks occur give the order of radial symmetry that the corresponding structures possess. Once the complete power spectra for a specimen have been retrieved, representative spectra are taken from each structural area of the coccolith. These data, though substantially condensed relative to the original image data, are still highly multivariate and this renders interpretation difficult. For this reason the data are input to a principal components analysis for further processing.

Principal components analysis

Principal components analysis converts a set of variables into a set of principal components. There are as many principal components as there are variables, but the first few principal components normally account for a larger proportion of the total variance than an equivalent combination of the original variables. Consequently, principal components are used to reduce the dimensionality of the data set whilst retaining most of the original variance. This form of analysis is general, makes no prior assumptions about the data, and tests no hypotheses, it simply transforms the data into another form which may be more useful for interpretation.

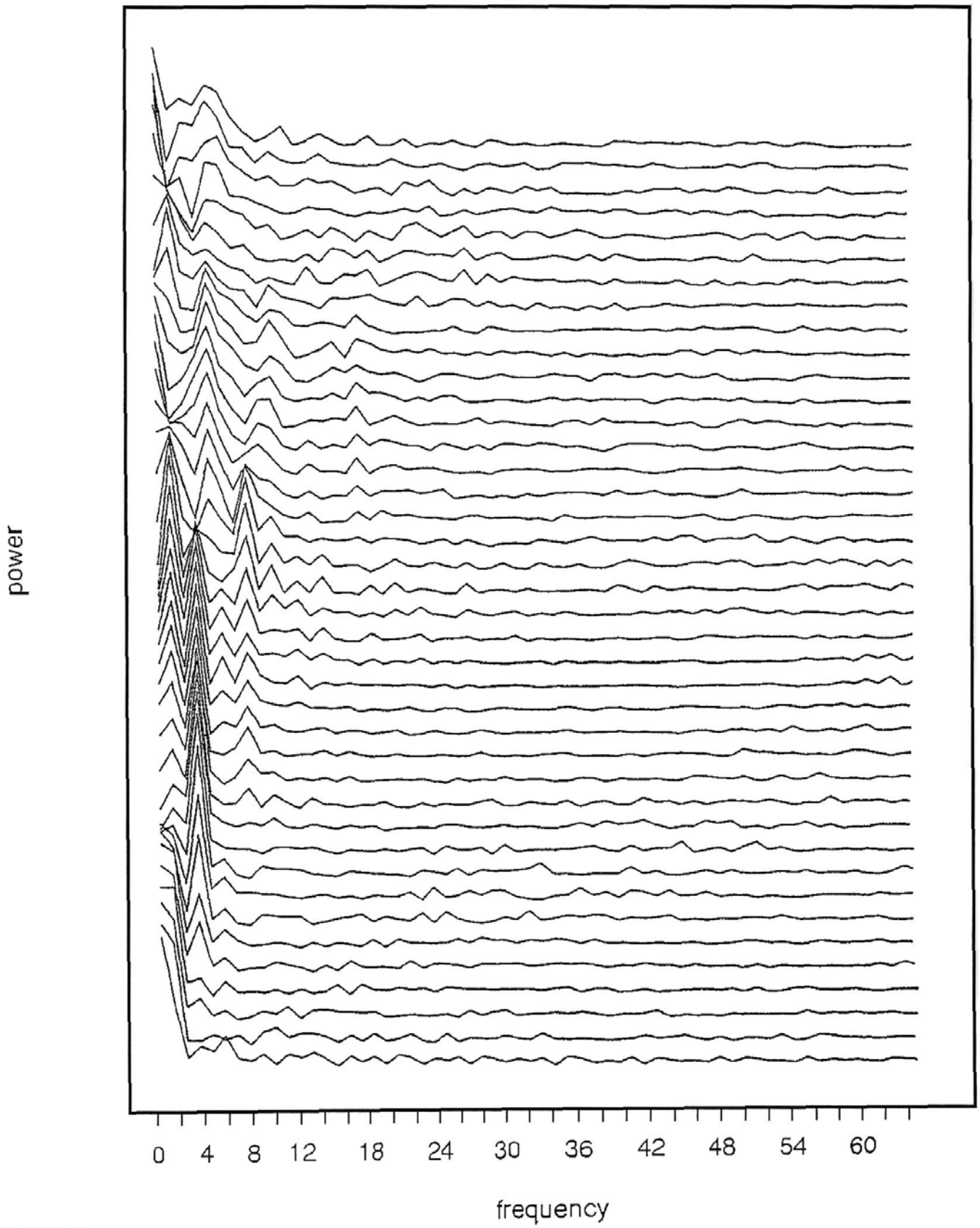
Principal components analysis uses the data variances and covariances to calculate eigenvectors and eigenvalues for each variable. Eigenvectors describe the direction of maximum variability (or spread) of a data cluster and are mutually orthogonal. This information can be displayed as the orientations of the principal axes of an m dimensional ellipsoid, where m represents the number of variables. The eigenvalues give the lengths of the successive principal semi-axes which represent the total variance of the data set and their lengths are proportional to the amount of the total variance represented by each variable. Thus the longest axis of the ellipse refers to the combination of variables with the strongest intercorrelations.

Each principal component can be directly related to its component variables by investigating the eigenvector loadings. There are numeric values for each original variable, which are defined as the coefficients of the linear equation that defines the eigenvector. By determining which variables have the largest absolute values on the loadings it is possible to discover which set of variables make the most significant contribution to each principal component. These can then be related to the structures that they represent.

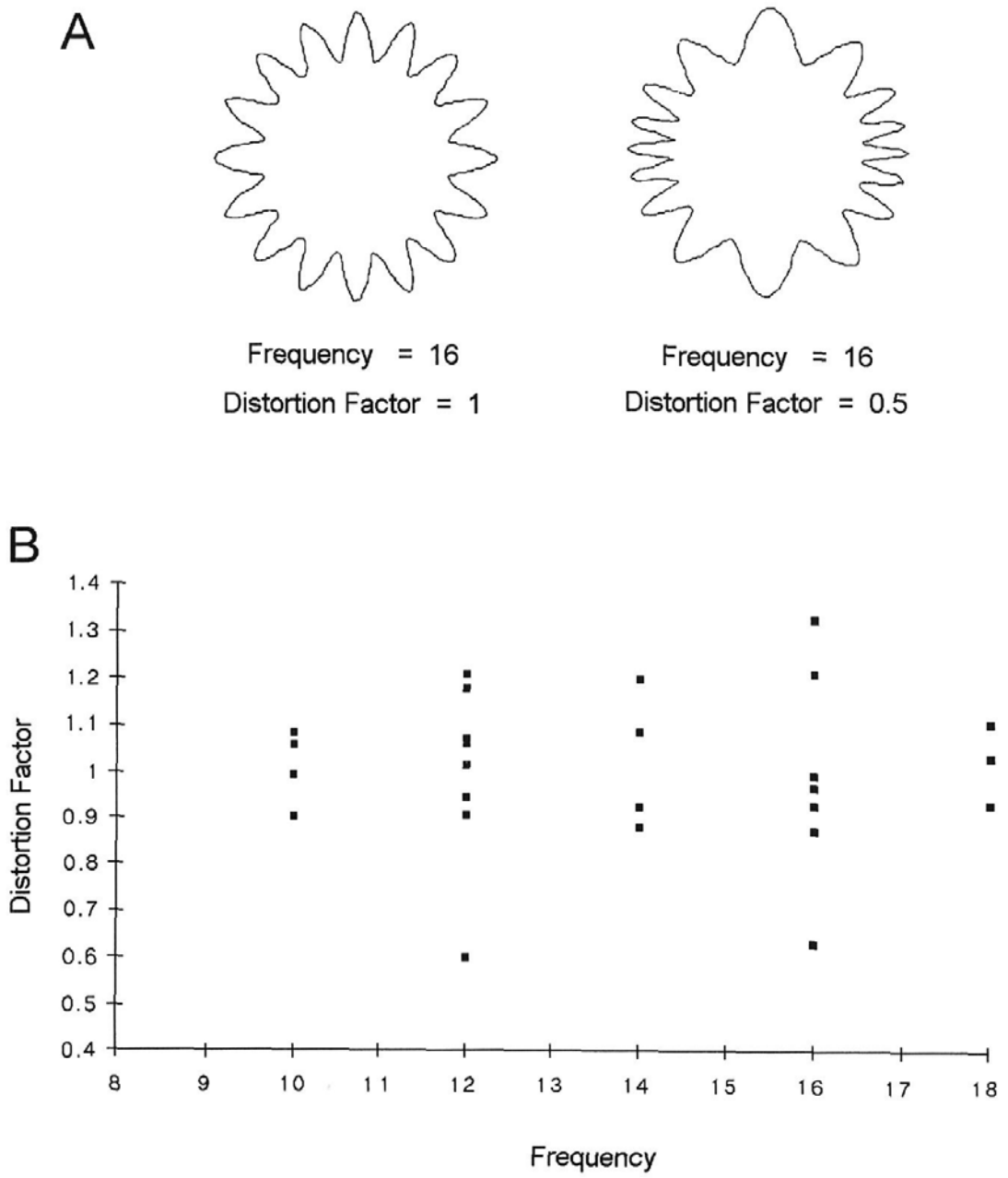
One of the most important uses of principal component analysis is to reduce the dimensionality of data by retaining only a small number of the strongest principal components. The multiple variables entering the principal components analysis in the present context are the powers at each frequency on a Fourier spectrum. Principal component analysis has been previously applied for this purpose on outline data by Delmet and Anstey (1974) and Foote (1989).

Discriminant function analysis

The process of discrimination finds a linear combination of variables which produce the maximum difference between predefined groups. A discriminant function allows the transforma-



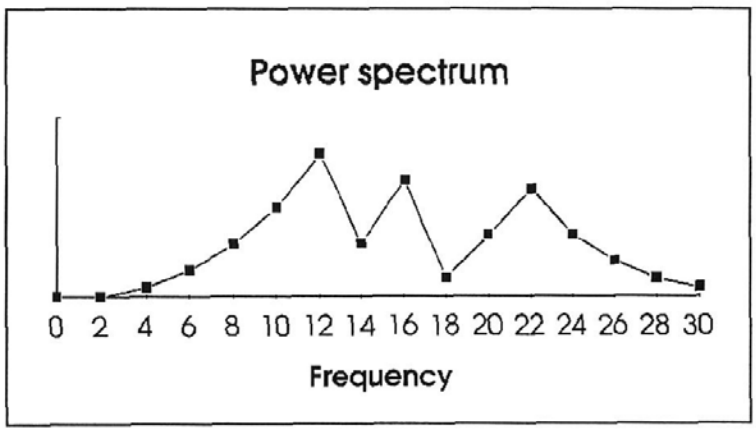
TEXT-FIGURE 6
The complete power spectra data retrieved from a specimen of *Prediscosphaera columnata*. The order of symmetry of the central area structure is represented by a strong peak at Frequency = 4 in the lower half of the plot. The order of symmetry of the rim area is represented by a weak peak at Frequency = 16 in the upper part of the plot.



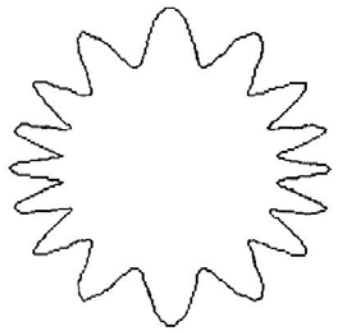
TEXT-FIGURE 7
Examples of simulated morphologies showing the effects of different distortion factors (Text-figure 7A) and the distribution of the complete set of simulated data items in frequency/distortion factor space (Text-figure 7B).

tion of multiple measurements of an object into a single discriminant score which represents its position along the discriminant function axis. Discriminant function analysis finds the transform which "gives the minimum ratio of the difference between a pair of group multivariate means to the multivariate variance within the groups" (Davis 1973).

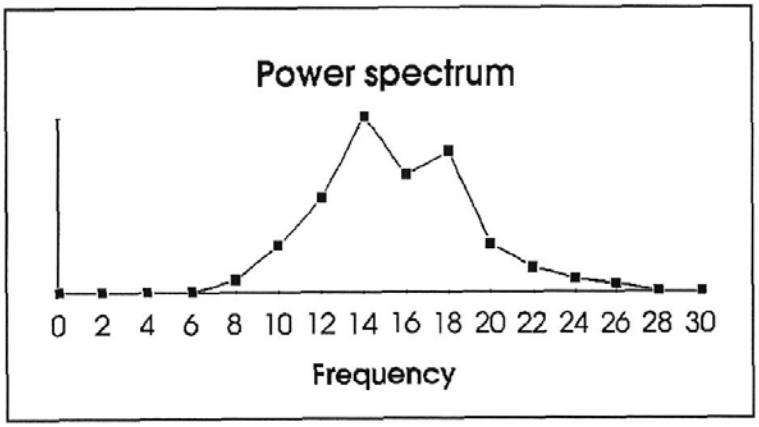
Computing a discriminant function for a data set involves two separate stages. Firstly a "training set" is taken from the original data, composed of selected specimens from two groups which are defined before the analysis begins. The user controls the discriminant function by the choice of training groups; any factor could, in principle, be discriminated with an appropriate choice of training groups. Discriminant function analysis is appropriate in the present context because we can select such groups to discriminate taxonomic, evolutionary and/or stratigraphic differences. There are no generally accepted guidelines for the size of training set; the larger the training set, the more reliable the results will be, but it may be less specific to the desired factor. In the present study the training groups were chosen to include 50% of the specimens and so that the discriminant function scores reflected morphological or taxonomical differences. The linear discriminant function is then calculated from this training set and is used to calculate discriminant function scores for each



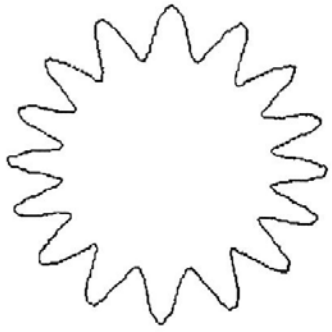
C



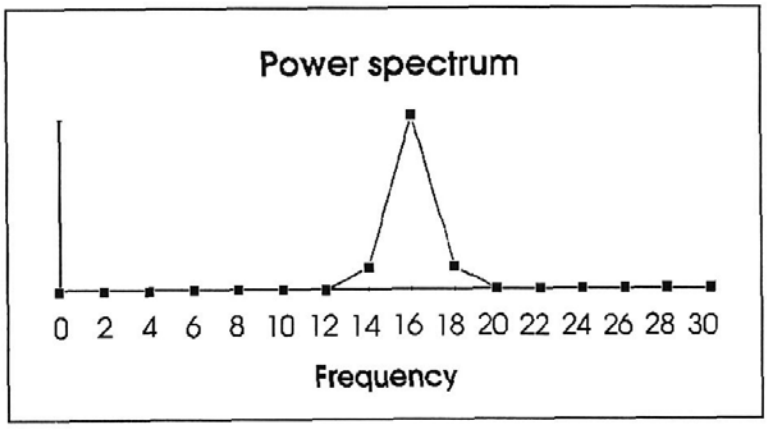
Distortion factor 0.62



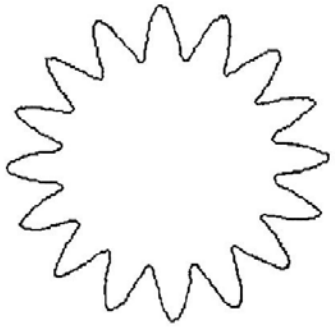
B



Distortion factor 1.21

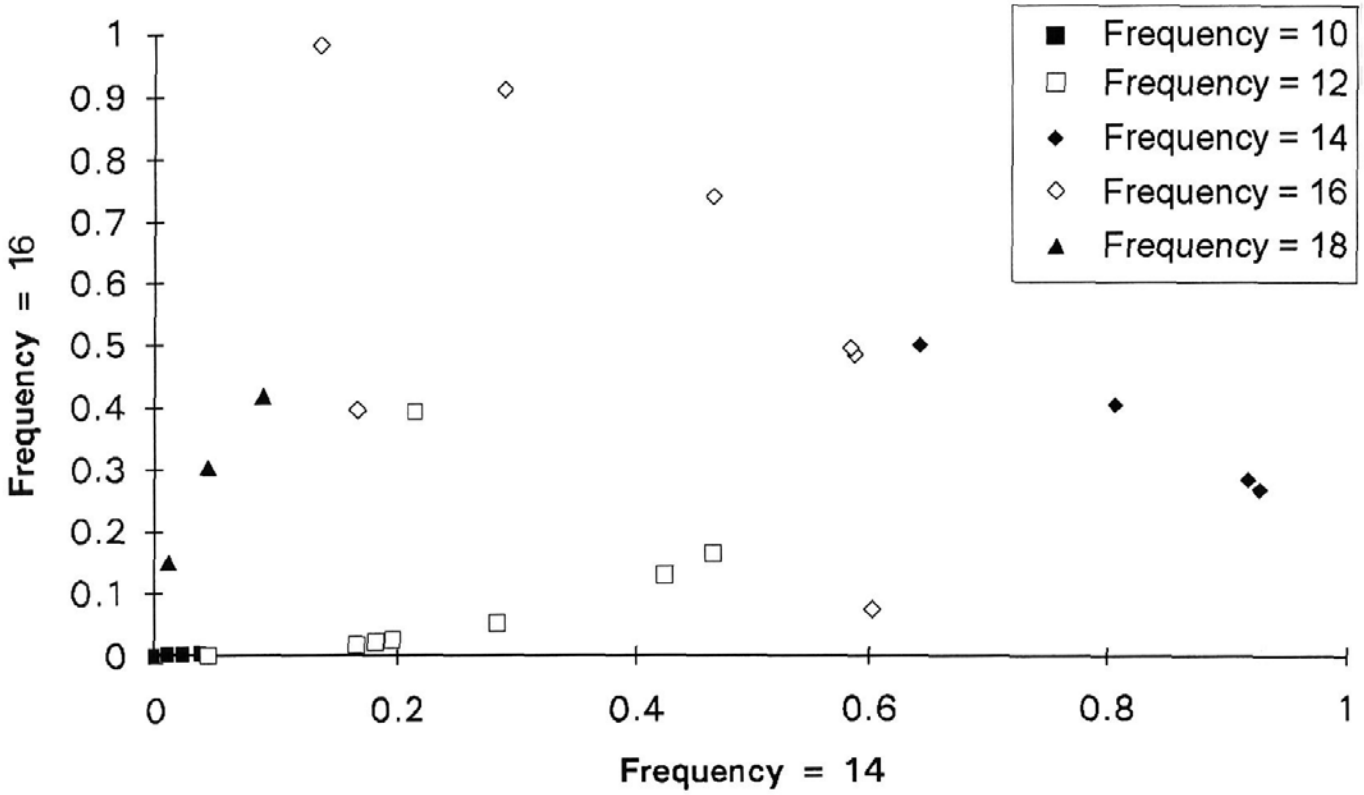


A



Distortion factor 0.96

TEXT-FIGURE 8
The effect on a power spectrum of increasing distortion factors which represent the interaction of different variables. The single sharp peak at a frequency of 16 (Text-figure 8A), is produced by relatively undistorted data. As the organisms become more distorted their power spectra reflect this (Text-figures 8B and 8C), and it becomes difficult to identify the base frequency.



TEXT-FIGURE 9
A bivariate scatter produced by ordinating the complete simulated data set against powers at frequencies 14 and 16 from Fourier transforms. The morphologies based on frequencies of 10 to 18 ("species") are not well separated.

group being analyzed. An initial measure of the efficiency of the technique is calculated by the number of training set items that are misclassified. A 90% efficiency rate is normally regarded as the minimum acceptable for this process, and all the real data sets used in the current analyses have efficiencies of 95%.

The second stage of the calculation involves using this discriminant function to predict the groups to which the non-training set data belong. A score for each data item is calculated from the discriminant function for the training set. The transformation used is:

1). Standardization of the entire data set using the combined training group means and standard deviations.

2). Matrix multiplication of the standardized data matrix by the principal component vectors calculated from the training set.

2). Matrix multiplication of the principal component scores matrix by the linear discriminant function to produce discriminant function scores for the entire data set.

Discriminant function analysis could not be applied directly to the Fourier output due to difficulties of inverting the variance-covariance matrix, resulting from high correlation between harmonics. Furthermore, it is not advisable to apply discriminant function analysis in a situation where the sample sizes involved are small relative to the number of Fourier-derived variables.

The nearness of the resultant discriminant function score to the central point of each training group can then be calculated, and the probability that the specimen belongs to either group can be determined.

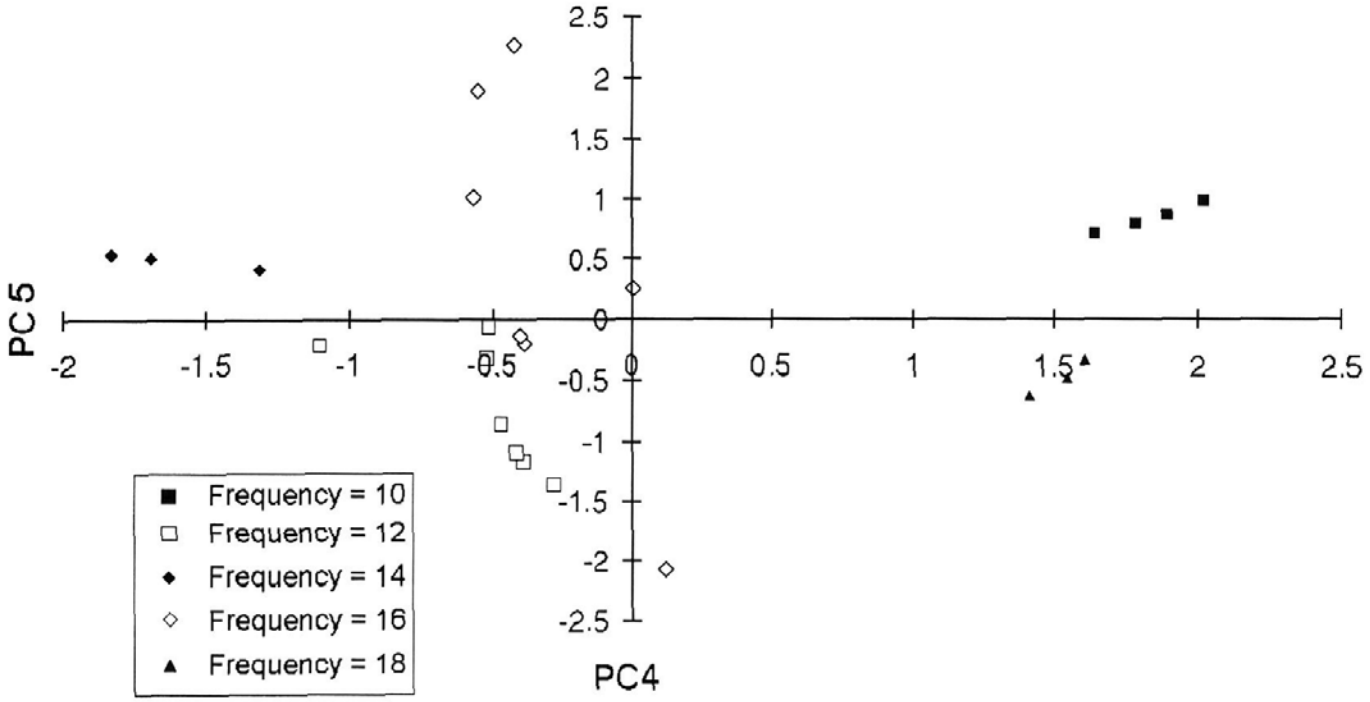
This process uses the same equation for the projection of all data points, allowing comparisons of how the taxa change through time to be made by plotting the discriminant function scores against stratigraphic position.

ASSESSING DATA COMPARABILITY, DESCRIPTIVE EFFICIENCY AND HOMOLOGY

Data retrieved from an image analysis system may be in several forms; for example direct measurements from an object, or as raster values that can represent colour or grey-level information for each pixel of a complete image. When direct measurements are used it is easy to ensure comparability of values between specimens and the assessment of homology is relatively straight-forward, but for a raster display such direct comparisons are problematic.

Raw data

Raster images consist of a matrix of pixels having varying values representing shades of grey (grey-levels). For any pixel value at a matrix position within an image there is a pixel value in the corresponding position in another image. These values are operationally comparable but certainly not biologically homologous: the position of the object within the pixel display



TEXT-FIGURE 10
Bivariate principal component (PC4 vs. PC5) plot for the simulated data set. The "species" are well discriminated.

will vary whilst the position of the pixel matrix is fixed. Futhermore, the pixel data are not efficient descriptors of the object: a very large number of pixels may be involved and the relationship with shape descriptors is indirect.

To allow valid comparisons to be made, a point of origin has to be determined which registers the position of the specimen. In the present case, the origin is the centre of symmetry of the coccolith and pixel values at the same relative position to this point in concentric scans may then be compared. Bookstein (1990) states that the use of regular angular increments in scans is "defensible only in the absence of all other information about homology", but in this case the scanning is an essential intermediate stage in the process of quantifying the homologous characters.

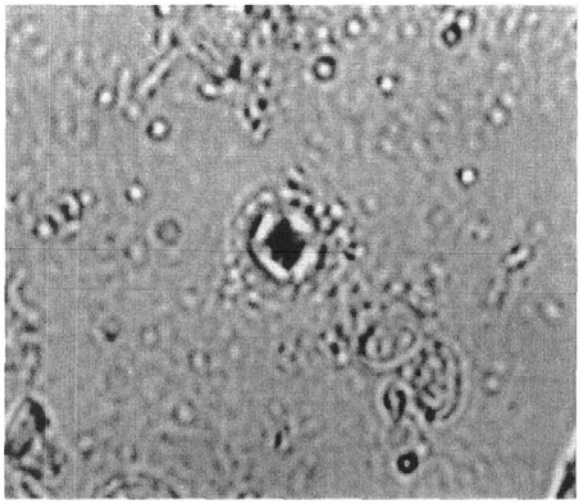
The raw data resulting from the scanning are all retrieved using the same process so mathematically comparability can be assumed.

Fourier power spectra

The interpretation of the Fourier output from the coccolith information is only concerned with the homologous characters "order of symmetry of the central area structure" and "number of plates in the rim". The mathematical comparability of the input data has been ascertained and, since the Fourier transform is applied to all the data in the same manner, the Fourier output can also be taken to be mathematically comparable.

The question of whether the power spectra are efficient shape descriptors or have any biological meaning is more complex. Bookstein (1990) argues that Fourier "features" have "no direct translation into the language of biological form-correspondance" but acknowledges that the technique may have a use in defining a morphospace in which forms can be ordinated and states: "Fourier coefficients may aid a numerical discrimination". Here, the aim is to assess the Fourier output in terms of its correlation with the pre-defined homologous structures of the original coccoliths: if the relationship is good, we can be justified in using the Fourier results in morphometric analysis.

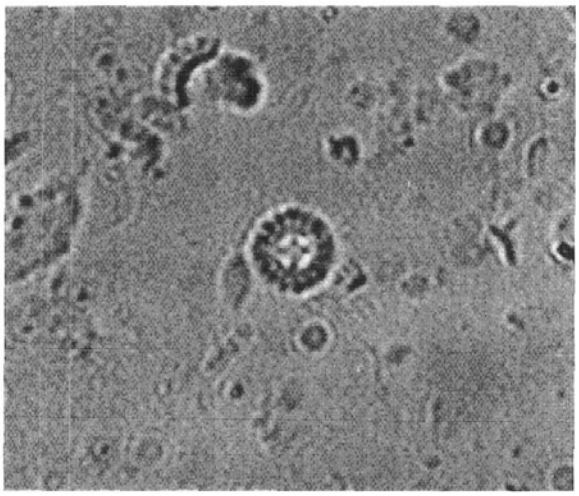
There are difficulties involved in relating individual Fourier frequencies to structural information: it is not normally possible to take a single peak from the power spectra and state that it definitively represents a homologous character because structural information may be spread over several peaks and, conversely, one peak may contain information from more than one structure. For example, in specimens of the genus *Prediscosphaera*, the part of the power spectra relating to the central region of the coccolith contains a peak at a frequency of F=4 (text-figure 6). This can be related to the central cross in this type of coccolith and represents the homologous character "order of symmetry of the central area structure." However the difficulty in interpreting these spectra is shown by the presence of other peaks at related frequencies, for example at a frequency of F=8. It is necessary to determine which of the peaks contain morphologically and biologically important information. In tests using the real data set the spread of structural information across the power spectra similarly meant that the Fourier process did not produce a statistically significant separation between different



a

10μm

b



TEXT-FIGURE 11
Photographs of specimens of *Prediscosphaera columnata* (Text-figure 11A) and *Prediscosphaera spinosa* (Text-figure 11B).

coccolith genera. It was therefore necessary to apply further processing.

Principal components

It is to be argued here that the use of principal components analysis applied to multivariate data (power at each frequency) from Fourier spectra improves the efficiency of shape description and enhances the correlation between morphometric descriptors and homologous attributes. This is possible because the dispersal of homologous structural information across the power spectrum causes degrees of intercorrelation between frequencies which can be detected and condensed as principal component axes. However, the validity of such analyses can not be ascertained by using the data from real images, because the "true" result is not known. To overcome this, a simulated control data set was created: this was based on a geometrical construction in which "homologous structures" could be defined. The simulated structures are representable as outlines but their properties are designed to be similar to those of the greylevel scans used here.

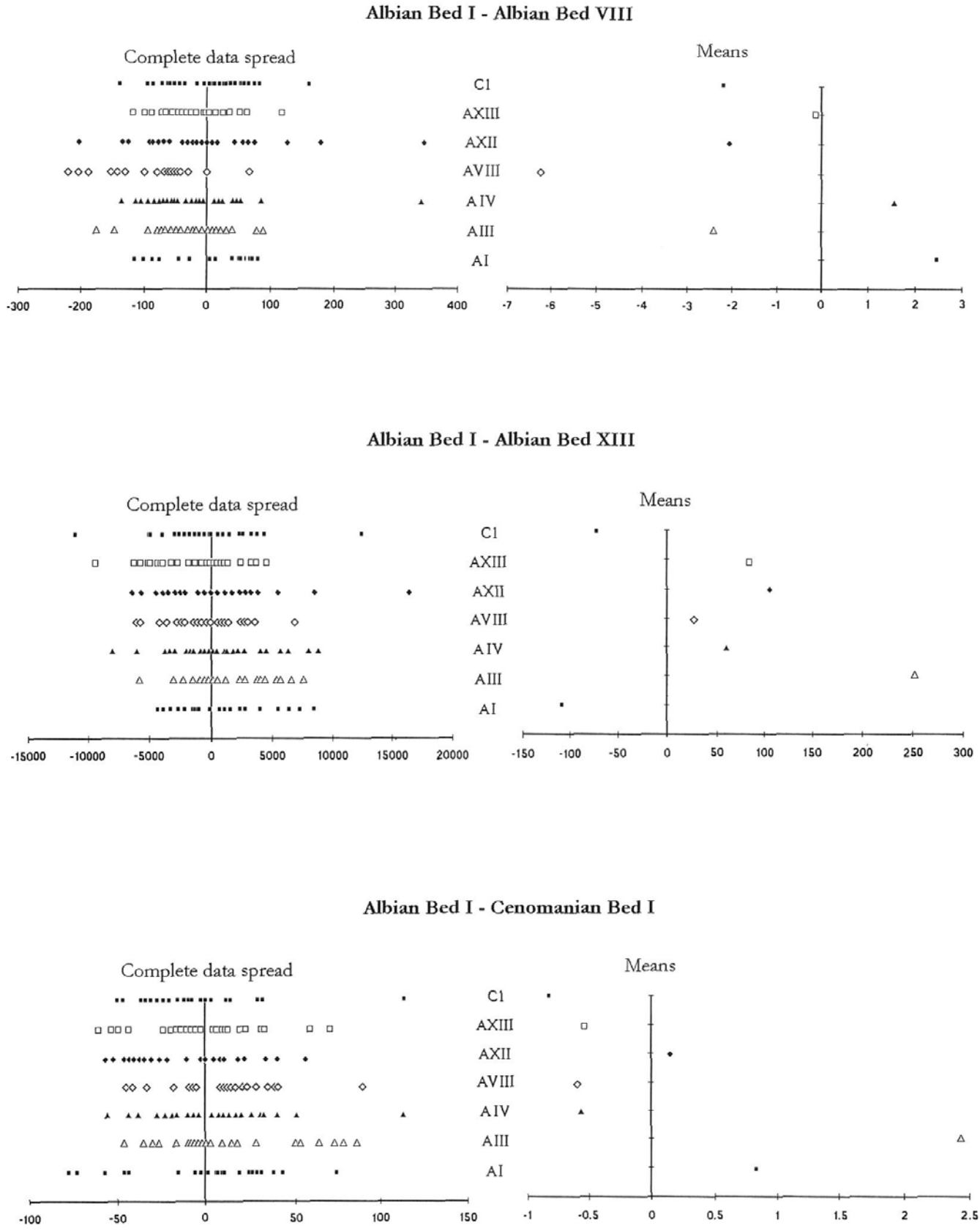
The simulated data set contains simple outlines consisting of regularly repeating cosine wave forms with frequencies of F=10, F=12, F=14, F=16 and F=18. These frequencies were deemed to be different states of the homologous character "number of outline peaks" and "species" were distinguished by the base frequency of the outline. These outlines were then systematically distorted by factors which had a random distribution about one. The nearer the distortion factor to one, the less distortion was produced. text-figure 7 shows examples of the distorted images and the distribution of the test data in frequency-distortion factor space. The nature of the distortion was designed to be analogous to the departure from perfect radial symmetry that accompanies ellipticity in coccoliths. More generally, it could be regarded as analogous in effect to non-cyclic distortions in outline shape. The data input to the Fourier analysis consist of measurements representing the radius of the image from a central point to a series of points on the image outline (comparable to grey-levels at specified circular or elliptical scans). A successful analysis would allow the identification of the different "species" by recovering the "homologous" character information at the frequencies of F=10, F=12, etc.

Since the real data produce power spectra that are distorted by the distribution of symmetry information among Fourier frequencies, the simulated data need to be investigated to determine the nature of this effect. text-figure 8 shows single power spectra produced by increasingly distorted images with a base frequency of 16. Sharp peaks on the power spectra are associated with small distortion factors.

If the simulated data are ordinated on axes for any two different frequencies, a bivariate plot such as that in text-figure 9 is obtained. Relatively undistorted specimens are well separated by the Fourier process and plot near a value of one on the corresponding axis, but as the distortion increases it becomes progressively more difficult to separate specimens from different species. There is no significant separation between a specimen with a base frequency of F=16, distortion factor 1.212 and that with a base frequency of F=14, distortion factor 1.192.

Principal components analysis was then applied to the Fourier spectral data of the simulated morphologies. The major principal components (PC1, PC2 and PC3) were found to correspond to linear combinations of variables which did not correlate strongly with "homologous" information. However, text-figure 10 shows the bivariate principal component scores plot for PC4 against PC5: this produces a good separation of the different "species", with five different groupings being visible. The separation is much improved over the corresponding Fourier plot (see text-figure 9), and was found to be statistically significant. The five different groups reflect the five different "species" created in the original data set.

These results show that "homologous" information was recovered by the principal components analysis. Suites of non-homologous, but correlated, characters in the power spectra



TEXT-FIGURE 12
Discriminant function score and mean score plots for the three different discriminant function analyses. Legend: AI = Albian Bed I, AIII = Albian Bed III, AIV = Albian Bed IV, AVIII = Albian Bed VIII, AXII =Albian Bed XII, AXIII = Albian Bed XIII, C1 = Cenomanian Bed 1. All show significant changes through time.

TABLE 1
The length width and elliptical ratio of the typical *Prediscosphaera* and *Zeugrhabdotus* specimens.

Specimen	Length	Width	Elliptical Ratio
Prediscosphaera	60	28	0.46
Zeugrhabdotus	119	60	0.50

combine to produce principal components which are better correlated with homologous characters. Indeed, each step in the transformation from outline data to power spectrum to principal components has been accompanied by an increase in correlation between morphological descriptors and the "homologous" attributes, and principal component analysis of outline power spectra can be justified as a methodology. In the case of our real image data, the transformation from the pixel array to elliptical scan to power spectrum to principal component can similarly be expected to improve correlation with homologous attributes. When applied to the real data, principal components analysis yielded axes that improved the taxonomic separation, but discriminant functions of the principal component scores were required to find oblique axes corresponding to the taxonomic properties of interest.

ANALYSIS OF COCCOLITH IMAGES

The discrimination of separate genera

The first aim of the case study was to determine whether the complete procedure could separate the coccolith genera *Prediscosphaera* and *Zeugrhabdotus* successfully. In this analysis the genera used were processed automatically, and no operator input was required once the original specimens were identified. The process includes the use of a discriminant function analysis to obtain the best possible separation of the data, and the significance of the results was checked using the Mann-Whitney test for equivalency of medians of two data groups.

Thirty specimens of each type (60 in total) were input to the analysis. This number was chosen because it fulfils the requirements of each separate analysis technique; for example the discriminant function can only process a maximum of twenty-five variables per specimen (due to limitations on the number of significant figures available in the processing package) but a valid analysis requires that many more specimens than variables must be processed. The time taken to retrieve each image places an upper limit on the number that can be used.

Fifteen specimens of each genus were used as the discriminant function training set, with the intention of training the analysis to discriminate between morphological features. The data derived from these specimens were processed by the Fourier and principal component analyses before being used to define the discriminant function. The remaining specimens were then subjected to the same procedure which culminated in principal component scores being produced for the two genera. 82% of the specimens are correctly separated by the procedure into genera, and the result is statistically significant.

This result shows that the automatic procedure is sensitive enough to produce viable results from real data, although the separation is not as accurate as could be achieved by a trained paleontologist studying the same data. This discrepancy is partly accounted for by the flexibility of a manual system. The automatic system only processes one image for a specimen, whereas a human operator can alter the focus and lighting conditions of an image and use crossed-polar as well as plane polarized light to achieve an identification.

The discrimination of separate species

To be generally useful for paleontological work this type of process must be able to discriminate between separate species at an acceptably accurate level. To test this coccolith images from two different *Prediscosphaera* species were retrieved. Text-figure 11A shows a specimen of the *P. columnata* species, and text-figure 11B shows an example of a *P. spinosa*. *P. columnata* has a rougOhly circular outline with a central X, whilst *P. spinosa* is more elliptical and the central cross is parallel to the axes of the ellipse. The specimens were processed using the same analysis method as for the different genera test, with thirty specimens being used as the training set for the discriminant function.

Seventy-nine percent of the specimens are correctly grouped by the analysis. This misclassification percentage is still high when compared to the results obtained by an experienced paleontologist. However, it is comparable to the results obtained in the different genera test even though the morphological differences of the two species are smaller than the differences between genera. The similarity in misclassification percentage suggests that the major limitations of the analysis procedure are caused by factors that are not related to the morphologies of the specimens, such as the resolution of the original images. If major changes in the preservation states or morphologies were affecting the data these would be expected to cause larger variations in the results.

Investigating morphological changes through time

The final test determined whether the process could discriminate changes occurring within a single genus through time. The recognition of such changes is important in applied paleontological research but these changes tend to be small since specimens with widely differing morphologies are likely to be classified as separate species.

The coccoliths in the study are taken from the Gault Clay and Cenomanian Chalk Marl formations at Copt Point and The Warren, near Folkestone in Kent. Samples of 30 coccoliths of the genus *Prediscosphaera* were taken from each of 5 Albian beds and the 2 lowest Cenomanian beds, giving 210 specimens from the 7 stratigraphic horizons. The classification systems used for these beds are shown in Tables 2 and 3 and the beds used are marked with an asterisk. These beds were chosen to give a series of samples that represent the complete section being studied.

Three separate analyses were run for the data set, each consisting of the Fourier analysis, principal component analysis and discriminant function analysis for three different pairs of stratigraphic horizons. Each discriminant function analysis produced a discriminant function coefficient for a training set which consisted of 15 specimens from each species, and which was then applied to the complete data set for all 7 stratigraphic horizons. The purpose of this procedure was to define an index of morphology or taxonomy calculated for large numbers of specimens from many stratigraphic horizons. The index used is the discriminant function of the principal component scores of standardized measurements and Fourier transform powers.

The following pairs of stratigraphic horizons were used by the analysis:

TABLE 2
A summary of the characteristics of the Gault Clay beds studied. (After Price 1874; Jukes-Brown and Hill 1900; Owen 1958, 1963, 1971).

Bed No.	Lithology	Thickness (ft)	Zone	Subzone
XIII*	Black glauconitic clay with septarian nodules	24′	*Stoliczkaia dispar*	*Stoliczkaia dispar* *Mortoniceras perinflatum* *Arrhaphoceras substuderi*
XII*	Dark grey clay with glauconitic and septarian phosphatic nodules	3′ 3″	*Mortoniceras inflatum*	*Mortoniceras altonense*
XI	Pale grey marl with greensand seam	56′ 3″		*Callihoplites auritus*
X	Hard pale grey marl	5′ 1″		*Hysteroceras varicosum*
IX	Pale grey marly clay	9′ 5″		*Hysteroceras orbignyi*
VIII*	Darker clay with two lines of nodules	0′ 10″	*Euhoplites lautus*	*Dipoloceras cristatum*
VII	Dark clay, highly fossiliferous	6′ 2″		*Anahoplites daviesi*
VI	Mottled dark clay with lighter markings	1′		*Euhoplites nitidus*
V	Dark clay with lighter spots	1′ 6″		*Euhoplites meandrinus*
IV*	Lighter, unmottled clay	0′ 4″	*Hoplites dentatus*	*Mojsisovicsia subdelaruei*
III*	Light fawn clay	4′ 6″		*Dimorphoplites niobe*
II	Very dark clay	4′ 3″		*Anahoplites intermedius*
I*	Dark clay, greensand, phosphatic and pyritic nodules	10′ 1″		*Hoplites spathi* *Lyelliceras lyelli* *Hoplites eodentatus*

1). Albian Bed I and Cenomanian Bed II, the base and top of the section used.

2). Albian Bed I and Albian Bed XIII, the base and top of the Albian beds.

3). Albian Bed I and Albian Bed VIII, the base of the Albian and the top of the Lower Gault. (Owen 1963, 1971).

These horizons were chosen to give a representative sample of the data from the complete section.

The principal component scores were then checked by using bivariate plots of the scores for PC1, PC2 and PC3 at each stratigraphic horizon. Analysis of the bivariate plots for all the horizons show that none of the most significant principal component scores change in a manner that consistently reflects the morphology of the specimens. This result confirms the hypothesis that morphological changes do not occur along the orthogonal axes which are identified by the principal component analysis to separate variables.

The lack of significant results following the principal component analysis meant that discriminant function analyses were required for each test. For each discriminant function analysis the efficiency of the separation was checked by recording the percentage of specimens that were misclassified. For the Albian Bed I -Cenomanian Bed II test 25% of the specimens were misclassified, 28% were misclassified in the Albian Bed I -Albian Bed XIII test and 10% in the Albian Bed I - Albian Bed VIII test.

The coefficients produced by each discriminant function analysis were then used to produce the discriminant function scores for the complete data set. Text-figure 12 shows plots of the discriminant function scores and mean positions of these scores for the *Prediscosphaera* data set for the three different discriminant function coefficients.

In all tests changes in morphology were identified between Albian Beds XII and XIII, other breaks in the morphology were not common to all three analyses. The significance of this break was then tested using a Kruskal-Wallis test. The test produces a significant result for the Albian Bed I -Cenomanian Bed II and Albian Bed I -Albian Bed VIII tests, but is not significant in the Albian Bed I - Albian Bed XIII test. The misclassification percentage for specimens is also high for this last result and may be affected by the preservation conditions of the specimens from Albian Bed XIII.

TABLE 3
A summary of the characteristics of the Cenomanian Chalk Marl beds studied. (After Kennedy 1969).

Bed No.	Nature	Thickness (meters)	Fossils
2*	Alterations of thin often nodular limestones and thick shaly marls. Both are blue-grey and extensively burrowed. Glauconite common at base.	>15	Rich fauna, many sponges in the limestones, large *Inoceramus crippsi,* possible *Mantelliceras saxbii* assemblage
1	Glauconitic marl subdivided into 8 beds. Top marked by a hard thin nodular limestone with sponges. Intensively burrowed	4.8	Fossils other than sponges are rare

Thirteen significant variables which contribute to the species separation are identified by all three tests, Table 4, but it is extremely difficult to relate these to existing morphological details. For example, length of major axis is readily identifiable, but significant variables also occur at frequencies 40 and 42 in the central area of the coccoliths. Examination of the *Prediscosphaera* specimens show no obvious corresponding structural elements. It is possible that these variables represent changing morphological features that have not previously been identified as important. To make full use of the potential of the results an index of significant variables from different coccolith taxa would have to be created.

These results show that changes within a genus through time can be identified, although it is difficult to quantify the differences in terms of existing recognized homologous characters.

DISCUSSION AND CONCLUSIONS

Each stage of the data analysis has been investigated to determine the mathematical comparability, efficiency of shape description and relationship with homologous attributes. As the analysis proceeds it becomes more difficult to relate the output to single input variables, but the results enhance the representation of morphological information. The final output shows the successful identification of the most important structural elements from the original images in a form which allows taxonomic comparisons to be made. This is possible because the principal components analysis and the discriminant function analysis increase the correlation with predefined homologous properties of the coccoliths, relative to the raw data and Fourier spectra.

The most difficult results to interpret in an homologous manner are those from the Fourier transform, because of the problem of information being spread between different Fourier frequencies. However, data referring to the original structural elements are contained within this output and are identified by the subsequent processing.

The principal components analysis succeeds in enhancing the Fourier output to a point where homologous characters can be recovered, but the separation of the specifically desired factors is produced by the use of the discriminant function analysis.

Automatic analysis procedures are sensitive to small variations within taxa, but the results are often difficult to interpret in terms of existing taxonomic schemes. For future development new indices of significant variables may have to be created that will allow these variables to be compared objectively and allow more rigorous comparisons of different taxa to be made.

ACKNOWLEDGMENTS

The computer equipment, specimens and partial funding for this research were provided by BP Research at Sunbury on Thames as part of a CASE studentship (NERC reference number GT4 88 GS 49).

REFERENCES

ANSTEY, R.L. and DELMET, D.A., 1973. Fourier analysis of zooecial shapes in fossil tubular bryozoans. Geological Society of America Bulletin 84: 1753-1764.

BENSON, R.H., 1967. Muscle-scar patterns of Pleistocene (Kansan) ostracodes. Pp 211-241. In Teichert, C. and Yochelson, E.L. (eds.) Essays in Palaeontology and Stratigraphy: Lawrence, Kansas, Kansas University Press.

BURKE, C.D., FULL, W.E., and GERNANT, R.E., 1987. Recognitiion of fossil freshwater ostracodes: Fourier shape analysis. Lethaia 20: 307-314.

CHRISTOPHER, R.A. and WATERS, J.A., 1974. Fourier series as a quantitive descriptor of miospore shape. Journal of Palaeontology 48: 697-709.

DAVIS, J.A., 1973. Statistics and data analysis in Geology, second edition. John Wiley and Sons, New York. 646 pp.

DELMET, D.A. and ANSTEY, R.L., 1974. Fourier analysis of morphological plasticity within an Ordovician bryozoan colony. Journal of Palaeontology 48(2): 217-226.

EHRLICH, R. and WEINBERG., B., 1970. An exact method for characterisation of grain shape. Journal of Sedimentary Petrology 40: 205-212.

ELDREDGE, N., 1979. Cladism and common sense. Pp 165-198 In Cracraft, J. and N. Eldredge, eds. Phylogenetic Analysis in Palaeontology. Columbia University Press. 233pp.

FERSON, S., ROHLF, F.J., and KOEHN, R.K., 1985. Measuring shape variation of two-dimensional outlines. Systematic Zoology, 34: 59-68.

FOOTE, M., 1989. Perimeter-based Fourier analysis: a new morphometric method applied to the trilobite cranium. Journal of Palaeonotology 63(6): 880-885.

FOOTE, M., 1991. Morphologic patterns of diversification: examples from trilobites. Palaeonotology 34(2): 461-485.

FULL, W.E. and EHRLICH., R., 1982. Some approaches for the location of centroids of Quartz grain outlines to increase homology between Fourier amplitude spectra. Mathematical Geology 14(1): 43-55.

FULL, W.E. and EHRLICH., R., 1986. Fundamental problems associated with "Eigenshape analysis" and similar "Factor" analysis procedures. *Mathematical Geology* 18(5): 451-463.

GARRATT, J.A. and SWAN, A.R.H., 1993. Morphological data from coccolith images. In Hamrsmid, B. and J.R. YOUNG. (eds.) Nannoplankton Research. I General Topics; Mesozoic Biostratigraphy. Knihovnicka Zemniho, MND Hodonin. 11-34.

HEALY-WILLIAMS, N., 1983. Fourier shape analysis of Globorotalia truncatulinoides from late Quaternary sediments in the southern Indian Ocean. Marine Micropalaeontology 8: 1-15.

HEALY-WILLIAMS, N., 1984. Qualitative image analysis: application to planktonic foraminiferal palaeoecology and evolution. Geobios, Memoire speciale 8: 425-432.

HEALY-WILLIAMS, N. and WILLIAMS, D.F., 1981. Fourier analysis of test shape of planktonic foraminifera. Nature 289: 485-487.

JUKES-BROWNE, A.J. and HILL, W., 1900. The Cretaceous rocks of Britain. Vol 1. Gault and Upper Greensand. Memoir of the Geological Survey, UK V 449pp.

KAESLER, R.L. and WATERS, J.A., 1972. Fourier analysis of the Ostracode margin. Geological Society of America Bulletin 83: 1167-1178.

KENNEDY, W.J., 1969. The correlation of the Lower Chalk of south-east England. Proceedings of the Geologists Association 80: 459-560.

LOHMANN, G.P. 1983: Eigenshape analysis of Microfossils: a general morphometric procedure for describing changes in shape. Mathematical Geology 15: 659-672.

MALMGREN, B.A., BERGGREN, W.A., and LOHMANN, G.P., 1984. Evidence for punctuated gradualism in the late neogene Globorotalia tumida lineage of planktonic foraminifera. Palaeobiology 9(4): 377-389.

OWEN, H.G., 1958. Lower Gault sections in the Northern Weald and the Zoning of the Lower Gault. Proceedings of the Geological Association, London 69: 148-165.

OWEN, H.G., 1963. Some sections in the Lower Gault of the Weald. Proceedings of the Geological Association, London 74: 35-53.

OWEN, H.G., 1971. Middle Albian stratigraphy in the Anglo-Paris basin. Bulletin of the British Museum (Natural History) Geology, London, Supplement 8:164pp.

PRICE, F.G.H., 1874. On the Gault of Folkestone. Quarterly Journal of the Geological Society 30: 342-368.

ROCK, N.M.S., 1988. Numerical Geology. Lecture notes in Earth Sciences. Springer-Verlag 18: 427pp.

TABLE 4
The significant variables identified in the *Prediscosphaera* test.

Significant variables
Length of major axis
f=10 c.a.
f=40 c.a.
f=42 c.a.
f=4 r.a.
f=6 r.a.
f=8 r.a.
f=14 r.a.
f=18 r.a.
f=22 r.a.
f=26 r.a.
f=56 r.a.

ROHLF, F.J., 1990. Fitting curves to outlines. 167-178. *In* Rohlf, F.J. and F.L. Bookstein, (eds.) *Proceedings of the Michigan Morphometrics Workshop*. Spec. Publ. No. 2, University of Michigan Museum of Zoology. 380pp.

SCHWARTZ, H.D. and SHANE, K.C., 1969. Measurement of particle shape by Fourier analysis. Sedimentology 13: 213-231.

SNEATH, P.H.A. and SOKAL, R.R., 1973. Numerical taxonomy. W.H. Freeman and Co. 573pp.

WAGNER, G.P, 1989. The origin or morphological characters and the biological basis of homology. Evolution. 43: 1157-1171.

WATERS, J.A. 1977: Quantification of shape by use of Fourier analysis: the Mississippian blastoid genus *Pentremites*. Palaeobiology 3: 288-299.

Manuscript received April 11, 1995
Manuscript accepted November 30, 1995