# The use of statistical classifiers for the discrimination of species of the genus *Gyrodactylus* (Monogenea) parasitizing salmonids

A. P. SHINN[1]*, J. W. KAY[2] *and* C. SOMMERVILLE[1]

[1] *Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, Scotland*
[2] *Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland*

(*Received 9 August 1999; revised 2 October 1999; accepted 3 October 1999*)

SUMMARY

This study applies flexible statistical methods to morphometric measurements obtained via light and scanning electron microscopy (SEM) to discriminate closely related species of *Gyrodactylus* parasitic on salmonids. For the first analysis, morphometric measurements taken from the opisthaptoral hooks and bars of 5 species of gyrodactylid were derived from images obtained by SEM and used to assess the prediction performance of 4 statistical methods (nearest neighbours; feed-forward neural network; projection pursuit regression and linear discriminant analysis). The performance of 2 methods, nearest neighbours and a feed-forward neural network provided perfect discrimination of *G. salaris* from 4 other species of *Gyrodactylus* when using measurements taken from only a single structure, the marginal hook. Data derived from images using light microscopy taken from the full complement of opisthaptoral hooks and bars were also tested and nearest neighbours and linear discriminant analysis gave perfect discrimination of *G. salaris* from *G. derjavini* Mikailov, 1975 and *G. truttae* Gläser, 1974. The nearest neighbours method had the least misclassifications and was therefore assessed further for the analysis of individual hooks. Five morphometric parameters from the marginal hook subset (total length, shaft length, sickle length, sickle proximal width and sickle distal width) gave near perfect discrimination of *G. salaris*. For perfect discrimination therefore, larger numbers of parameters are required at the light level than at the SEM level.

Key words: *Gyrodactylus salaris*, statistical classifiers, nearest neighbours, feed-forward neural network, projection pursuit regression, linear discriminant analysis.

## INTRODUCTION

The identification of many parasites relies heavily upon the comparison of their morphological and morphometric characters with other species of their respective genera. These characters may be attachment hooks or parts of an endo/exo-skeleton or organs which, when stained, have a characteristic shape. Many inadequacies in traditional methods of identification have been exposed, especially where pathogens require to be distinguished from closely related, but non-pathogenic, forms and where accurate monitoring of introduced parasite species is necessary. Recent evidence has demonstrated that translocation of fish across country borders has increased the rate of introduction of exotic parasite species to indigenous fish stocks with serious economic consequences (Kennedy, 1993). The recent introduction of several exotic metazoan parasites into the UK as documented by Gibson (1993) and Kennedy (1993), gives cause for concern. Ten of these parasite species are already established; for example, despite legal proscription of movements of infected fish, *Khawia sinensis*, a caryophyllaeid tapeworm, is spreading through Britain (Yeomans, Chubb & Sweeting, 1997). Further, the recent introduction of other serious waterborne diseases such as crayfish plague (Alderman, 1996) into the UK demonstrate the ability of infectious agents to translocate as a result of commercial activity. Some of the introduced parasites are known to be serious pathogens and their effect may be critical for conservation and fisheries management as well as aquaculture. For example, *Gyrodactylus* spp. are common ectoparasitic monogeneans with 400 species being described (Harris, 1993). One member of the genus, *Gyrodactylus salaris* Malmberg, 1957, is considered to be very highly pathogenic to some stocks of Atlantic salmon, *Salmo salar*, whereas other species of *Gyrodactylus* infecting salmonids have a generally low pathogenicity. *Gyrodactylus salaris* is responsible for the catastrophic decline in salmon stocks in Norway and has been demonstrated to be widespread in Norwegian rivers (Johnsen & Jensen, 1988; Mo, 1994) with a projected reduction in returning salmon of 20% (ca. 300 tons) (Johnsen & Jensen, 1986). *G. salaris* is now known from 10 neighbouring European countries, most recently Portugal and France (Johnston *et al.* 1996). To prevent its entry into the UK, *G. salaris* was made a

* Corresponding author: Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, Scotland. Tel: +44 1786 473171. Fax: +44 1786 472133. E-mail: aps1@stir.ac.uk

 

notifiable disease in 1988 under the 1937 and 1983 Diseases of Fish Acts of the UK. Studies of sample sites in England, Scotland and Wales by Shinn, Sommerville & Gibson (1995) and in Northern Ireland by Platten, McLoughlin & Shinn (1994) have shown that *G. salaris* is so far absent from the UK. If the UK's *G. salaris*-free status is to be maintained, it is essential to have in place accurate techniques to discriminate this pathogenic species from the other gyrodactylids infecting salmonids.

There is an urgent need to develop reliable methodologies that can confidently identify pathogens such as *G. salaris* that can be used by non-specialists e.g. river biologists and fish health diagnosticians. Specialists in monogenean taxonomy are few in number, and the volume of samples now generated from the regular screening of fish under the UK's health certification for *G. salaris* is too great. Mixed infections of *Gyrodactylus* do occur on salmonids (Shinn *et al*. 1995), host specificity cannot be assumed, and results from diagnostic screening are required quickly so that, if necessary, containment of an identified infection can be implemented rapidly. Thus, the benefits of an automated system that can reliably identify *G. salaris* in samples are clear. Advances in molecular biological techniques, such as species-specific probes are under development (Cunningham *et al*. 1995*a*, *b*); however, at present their implementation is expensive, time consuming and requires a high degree of expertize. The use of statistical classifiers and such technologies as artificial feed-forward neural networks (FFNN) (Kay, Shinn & Sommerville, 1999), present a rapid reliable alternative to traditional methods. Once the statistical classification system has been trained and validated, it can be disseminated widely among potential users for whom the technique will be straight-forward to perform and will give clear results. The development of automated systems which can be used widely by non-specialists will allow for rapid, early detection at source and prevent translocation of potential pathogens.

## MATERIALS AND METHODS

### *Parasite collection*

A total of 88 sites in the UK with salmonid populations (*Salmo salar*, *S. trutta*, *Oncorhynchus mykiss* and *Salvelinus alpinus*) were sampled for *Gyrodactylus* specimens during the period May 1990 to April 1992 for studies using both light and scanning electron microscopy (SEM). Details of the sites sampled and the reference material used from national collections for validation are given in Table 1. For this study, samples were collected from wild and farmed salmonids in Ireland and the UK over a wide geographical range. All seasons were represented and environmental data recorded. *G. salaris* material was collected from Norway and Sweden. At selected sites, continuous sampling throughout 4 seasons was undertaken. The wide sampling programme was undertaken to ensure that variation due to host, locality, season and environmental conditions were included. Such a robust data set was considered necessary to have complete confidence in the techniques used and the result obtained.

### *Sample preparation and morphometric measurements*

For light microscopy, gyrodactylid specimens ($n = 470$ specimens selected out of a total of 648) were collected live from fish, where possible, or from those fixed in 80% alcohol. Regions of unconsolidated hook material such as the hamulus root portion are subject to distortion under fixation with alcohol and thus this measurement was not included within the morphometric measurements made. Morphometric measurements of the opisthaptoral hooks and bars were made from slide preparations of *Gyrodactylus* mounted in ammonium picrate glycerin (Malmberg, 1957) using an eye-piece graticule at $\times 100$, oil immersion lens on a BH2 Olympus binocular microscope with phase contrast illumination. Ten point to point morphometric measurements were made using the light microscope, 3 from the hamulus (total length, shaft length and point length) (Fig. 1C(a–c)), 2 from the ventral bar (total length and total width) (Fig. 1D(d–e)) and 5 from the marginal hook (total length, shaft length, sickle length, sickle proximal width and sickle distal width) (Fig. 1E(f–j)). To ensure continuity between samples, marginal hook number 8 was measured on each sample. Where this was not possible marginal hook number 7 was measured.

Marginal hooks analysed from scanning electron micrographs ($n = 222$) also shown in Fig. 1E(f–l), were processed and extracted following the procedures of Shinn, Gibson & Sommerville (1993) and the morphometric measurements made using those given by Shinn *et al*. (1996). A total of 7 point to point measurements were used in the analysis of SEM samples for statistical classification. In addition to the 5 measurements used for specimens prepared and measured using the light microscope, 2 additional measurements were made from SEM micrographs, the marginal hook sickle aperture and marginal hook sickle toe length were also used (Fig. 1E(k–l)). In contrast to the situation in light microscopy, the exact position of the marginal hook is lost following the hook extraction technique.

The figures given in Table 1 for each *Gyrodactylus* region sampled, represent the number of specimens used within this study and do not represent the total number of specimens collected from that particular site. For each site sampled, 10 specimens were selected randomly for measurement from all the gyrodactylids collected from all the hosts sampled for that particular site.

Table 1. Details of the *Gyrodactylus* spp. sample collection sites (05/1988–04/1992) and reference material used for the various methods of statistical classification

(The number of *Gyrodactylus* individuals collected from each respective host from each geographical location are given. Figures given represent the number of specimens of *Gyrodactylus* measured using the light microscope. Figures in parentheses denote samples of *Gyrodactylus* analysed using the SEM.)

| Location/host | No. of sites sampled | No. of gyrodactylids measured |
|---|---|---|
| Scotland | | |
| *O. mykiss* | 9 | 67 (20) |
| *S. salar* | 28 | 192 (30) |
| *S. trutta* | 14 | 95 (20) |
| Wales | | |
| *S. salar* | 14 | 59 (10) |
| *S. trutta* | 7 | 38 (10) |
| England | | |
| *O. mykiss* | 2 | 6 (—) |
| *S. salar* | 6 | 39 (—) |
| *S. trutta* | 5 | 26 (—) |
| *S. alpinus* | 1 | 10 (—) |
| Ireland | | |
| *S. salar* | 2 | 6 (—) |
| Norway | | |
| *S. salar*[1] | 1 | 10 (6) |
| Sweden | | |
| *S. salar*[2] | 5 | 75 (96) |
| *S. trutta*[2] | 1 | 10 (10) |
| Reference material | | |
| *G. colemanensis*[3] | 1 | — (10) |
| *G. truttae*[4] | 1 | 5 (—) |
| Sites and species sampled seasonally | | |
| *G. derjavini* | *O. mykiss* (Loch Awe, Scotland) (12/1989–04/1992) | |
| *G. caledoniensis* | *S. salar* (R. Allan, Scotland) (05/1990–04/1992) | |
| *G. derjavini* | *S. salar* (R. Allan, Scotland) (05/1990–04/1992) | |
| *G. truttae* | *S. trutta* (L. Airthrey, Scotland) (05/1990–04/1992) | |
| *G. salaris* | *S. salar* (R. Högvadsån[2], Sweden) (05/1991–03/1992) | |

[1] Collected by Dr T. A. Bakke.
[2] Collected by Dr G. Malmberg.
[3] Collected by Dr D. Cone.
[4] Collected by Dr P. D. Harris.

*Statistical classifiers*

The morphometric data from the gyrodactylid hooks and bars were used to assess the prediction performance of a number of statistical classification methods. Four methods were used, namely, nearest neighbours (NN), a feed-forward neural network (FFNN), project pursuit regression (PPR) and linear discriminant analysis (LDA); see McLachlan (1992), Haykin (1994) and Venables & Ripley (1997). The methods were implemented using the PC software S-PLUS 4 (1997) statistical package and software provided by Venables & Ripley (1997) to aid S-PLUS 4 users, although the statistical classification techniques can be conducted on a wide range of other packages available commercially or via free shareware. Linear discriminant analysis is a standard method. The other 3 methods are more complex and, in particular, they allow the fitting of non-linear boundaries between the classes. Certain complexity parameters must be determined to control the extent to which boundaries between classes are non-linear. For example, in the nearest neighbours (NN) method it is necessary to choose the number of neighbours. This is a complexity parameter: using only 1 nearest neighbour can result in very non-linear (more complex) boundaries between the classes in morphometric space whereas using, say, 9 nearest neighbours results in smoother boundaries.

The classification of specimens by linear discriminants (LDA), project pursuit regression (PPR), nearest neighbours (NN) and feed-forward neural networks (FFNN) has been discussed in more detail by Kay *et al.* (1999).

*Training the classifier and evaluating the test method*

A statistical classifier is trained by making use of previously classified data (e.g. morphometric data) and adjusting its parameters until the best possible

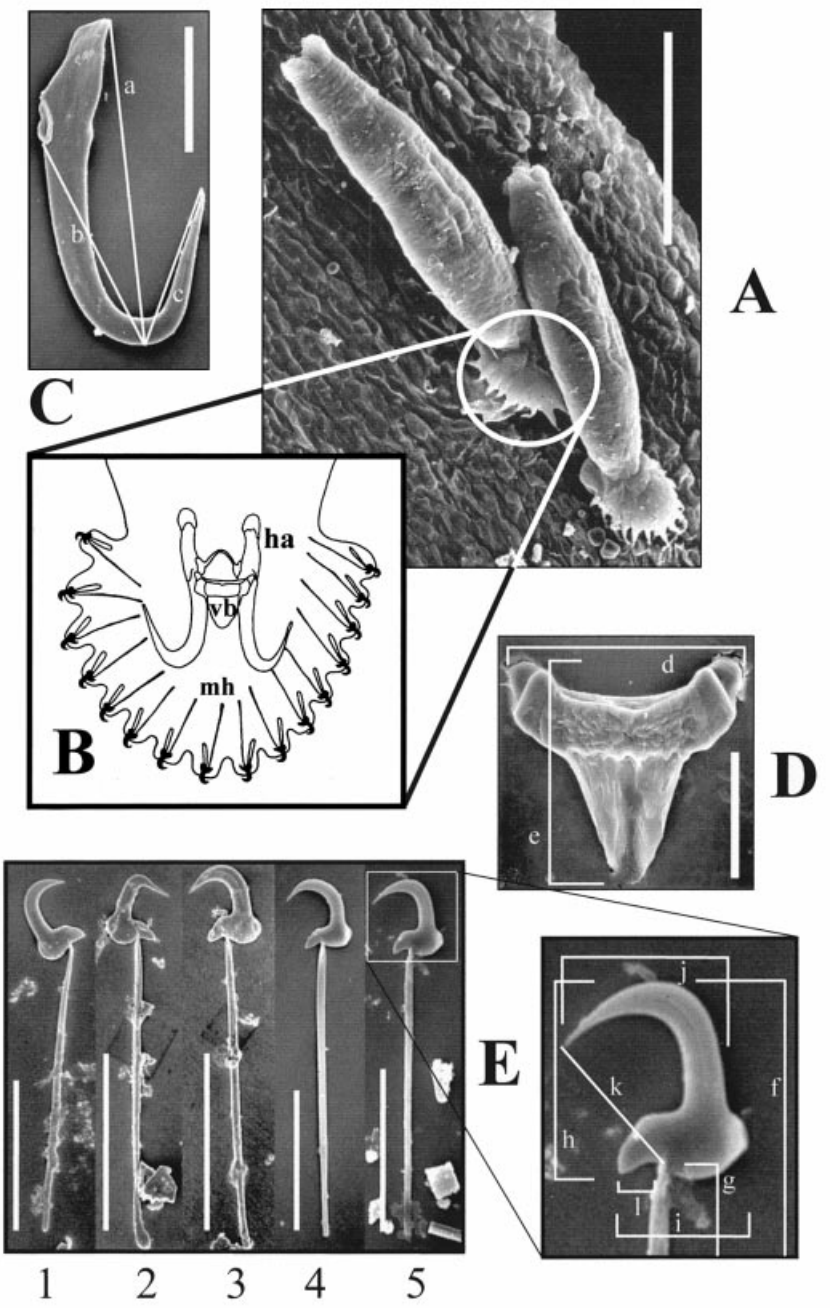


Fig. 1. Organization of the opisthaptoral hooks and bars of *Gyrodactylus* sp. and the presentation of the morphometric measurements derived from light and scanning electron microscope studies. (A) *Gyrodactylus* sp. attached to the epidermis of its host. (B) Diagrammatic representation of the arrangement of the opisthaptoral hooks and bars; ha, hamulus; vb, ventral bar; mh, marginal hook. (C) SEM of a hamulus extracted by sonication from *G. salaris* from *S. salar*: a, hamulus total length; b, hamulus shaft length; c, hamulus point length. (D) SEM of a sonication-released ventral bar of *G. derjavini* from *Oncorhyncus mykiss*: d, ventral bar length; e, ventral bar width. (E) marginal hooks of *Gyrodactylus* spp. released by proteolytic digestion: 1, *G. caledoniensis*; 2, *G. colemanensis*; 3, *G. derjavini*; 4, *G. salaris*; 5, *G. truttae*: f, marginal hook total length; g, marginal hook shaft length; h, marginal hook sickle length; i, marginal hook sickle proximal width; j, marginal hook sickle distal width; k, marginal hook sickle aperture; l, marginal hook sickle toe length. Measurements f–j were used for studies made using the light microscope and measurements f–l were made using the scanning electron microscope. Scale bars: A, 120 $\mu$m; C, 25 $\mu$m; D, 13·6 $\mu$m; E, 15 $\mu$m (scale bars all 15 $\mu$m).

classification accuracy is obtained. The trained classifier is then validated by applying it to test data (i.e. data from the same population as the training data) and then assessing the classification accuracy.

The method of stratified 3-fold cross-validation was used to assess the generalization error likely to be obtained when one of the classifiers is applied to new specimens. In this approach the available specimens are split randomly into 3 groups in proportion to the numbers of each type that are

available. One of the 3 groups of data is held back to form a test set, and the remaining data are used to build the classifier. The resulting rule is then applied to the test set and the predictions obtained for each test specimen are compared with the true types; thus the number of misclassifications is calculated. This procedure is repeated taking, in turn, each of the other 2 groups to be the test set, and the numbers of misclassifications are combined to form an overall estimate of the misclassification rate (EMR). This method makes efficient use of the available data.

A full explanation of the statistical classification methods selected, training and validating the classifier have been given by Kay *et al*. (1999).

## RESULTS

### *Statistical analysis of data from SEM-derived images of the marginal hook*

The identification of each specimen (training set + test set) for each of the 5 *Gyrodactylus* species investigated using SEM studies on only a single structure, the marginal hook, is given in Table 2. The data presented as misclassification matrices are based on 7 morphometric measurements. In the nearest neighbours (NN) method, 9 neighbours were used, 9 hidden units and a weight decay of 0·01 were employed in the FFNN classifier and 9 non-linear projections were used for the PPR method. As can be seen from Table 2, two of the statistical classifiers, namely nearest-neighbours and the feed-forward neural network, accurately discriminated all specimens of *G. salaris* from the other gyrodactylid species. This demonstrates that it is possible to distinguish *G. salaris* correctly from the other salmonid gyrodactylids studied here using morphometric data from only the marginal hook. The discrimination of the other *Gyrodactylus* species was not complete, for example, of the 10 specimens of *G. colemanensis* identified by the gyrodactylid taxonomist, all were correctly identified as *G. colemanensis* by the nearest neighbours classifier, however, of the 69 specimens identified by the gyrodactylid taxonomist as *G. derjavini*, the nearest neighbours correctly classified 62 of these as belonging to *G. derjavini* but misclassified 4 specimens as *G. colemanensis* (3 specimens from *O. mykiss* from Loch Butterstone and 1 specimen from *S. salar* from the River Nith), and 3 as *G. caledoniensis* (1 specimen each from *S. trutta* in the River Dalälven and *O. mykiss* in Loch Butterstone and Loch Awe).

### *Statistical analysis of data from light microscopy-derived images of all hooks and bars*

Analysis of gyrodactylid hooks and bars at the SEM level allowed for the perfect discrimination of *G. salaris* from other species of the genus parasitizing salmonids. However, anomalies are often generated when dealing with less clear images, a restriction imposed by the optical limitations of the microscope. In order to have complete confidence in the classifiers it was necessary to perform the analysis at the highest platform of resolution first (SEM), before attempting to analyse images obtained using the light microscope. Therefore, the statistical analyses were repeated on morphometric data from light microscope studies. The use of the light microscope would allow for a decrease in the specimen processing time whilst maintaining confidence in the discriminating ability of the classifier.

As the light data might perform less well, the initial analysis was based on all available measurements. Thus, 10 morphometric measurements using light microscopy were used. The same parameters used to run the statistical classifiers for the SEM data were also used here. The results are shown in Table 3.

The nearest neighbours classifier using 9 nearest neighbours gave the lowest estimates of generalization error and gave a perfect discrimination of *G. salaris* from *G. derjavini* and *G. truttae*. Similarly, the linear discriminant analysis gave perfect separation of *G. salaris* specimens from the other species studied, but the discrimination of *G. derjavini* from *G. truttae* was quite poor (EMR = 13·4 %). The other classifiers, FFNN and PPR, performed less well, misclassifying 1 and 2 specimens of *G. salaris* respectively as *G. derjavini* and *G. truttae*.

### *Analysis of data from individual structures derived from light microscopy*

The success of the nearest neighbours classifier to accurately discriminate *G. salaris* from the other salmonid gyrodactylids at the light microscope level, was further tested for its ability to classify single structures. Three structures, the hamulus, the marginal hook and the ventral bar from 3 species of *Gyrodactylus*, *G. salaris*, *G. derjavini* and *G. truttae*, were tested and the results are presented in Table 4. Two morphometric measurements were used to describe the ventral bar and, as can be seen from Table 4, this structure had an overall error rate of 21·7 %, of which 6·2 % involved misclassification of *G. salaris* and was not considered to be useful for the discrimination of these 3 gyrodactylids. The overall error rates for the hamulus and marginal hook were lower at 8·3 % and 10·9 % respectively, but both had misclassifications involving *G. salaris*. Using nearest neighbours therefore, it was possible to discriminate most specimens of *G. salaris* from *G. derjavini* and *G. truttae* using light microscope-derived data from only a single structure, either the hamulus (2 specimens of *G. salaris* misclassified as other gyrodactylid species) or marginal hook (1 specimen of *G. derjavini* misclassified as *G. salaris*). However, perfect discrimination was only achieved when using

Table 2. The application of 4 statistical classification methods to morphometric data derived from SEM studies of the marginal hook of 5 *Gyrodactylus* species

(Abbreviations: *Col*, *G. colemanensis*; *Derj*, *G. derjavini*; *Cal*, *G. caledoniensis*; *Trut*, *G. truttae*; *Sal*, *G. salaris*; EMR, Estimated Misclassification Rate.)

| | True class | | | | |
|---|---|---|---|---|---|
| Predicted class | *Col* | *Derj* | *Cal* | *Trut* | *Sal* |
| (A) Nearest Neighbours (NN) | | | | | |
| *Col* | 10 | 4 | 1 | 0 | 0 |
| *Derj* | 0 | 62 | 13 | 8 | 0 |
| *Cal* | 0 | 3 | 5 | 0 | 0 |
| *Trut* | 0 | 0 | 0 | 14 | 0 |
| *Sal* | 0 | 0 | 0 | 0 | 102 |
| EMR = 13·1 % | | | | | |
| (B) Feed-Forward Neural Network (FFNN) | | | | | |
| *Col* | 8 | 3 | 0 | 0 | 0 |
| *Derj* | 0 | 56 | 11 | 5 | 0 |
| *Cal* | 2 | 6 | 8 | 2 | 0 |
| *Trut* | 0 | 4 | 0 | 15 | 0 |
| *Sal* | 0 | 0 | 0 | 0 | 102 |
| EMR = 14·9 % | | | | | |
| (C) Projection Pursuit Regression (PPR) | | | | | |
| *Col* | 6 | 1 | 1 | 0 | 0 |
| *Derj* | 3 | 55 | 10 | 10 | 0 |
| *Cal* | 1 | 4 | 6 | 0 | 1 |
| *Trut* | 0 | 9 | 2 | 11 | 0 |
| *Sal* | 0 | 0 | 0 | 1 | 101 |
| EMR = 17.6 % | | | | | |
| (D) Linear Discriminant Analysis | | | | | |
| *Col* | 0 | 6 | 2 | 0 | 0 |
| *Derj* | 10 | 61 | 16 | 13 | 1 |
| *Cal* | 0 | 0 | 0 | 0 | 2 |
| *Trut* | 0 | 2 | 1 | 9 | 0 |
| *Sal* | 0 | 0 | 0 | 0 | 99 |
| EMR = 23.9 % | | | | | |

a hamulus, the ventral bar and a marginal hook together.

## DISCUSSION

Statistical classifiers were applied to simple point to point morphometric data taken from images of hard skeletal features and successfully discriminated several species of *Gyrodactylus* parasitizing salmonids which are otherwise difficult to separate. Host and environmental parameters such as temperature and salinity, have been shown to influence the morphological variation observed in gyrodactylid marginal hooks (Malmberg, 1970; Mo, 1991*a*, *b*, *c*). Specimens of *Gyrodactylus* forming the data set used for training the statistical classifiers within this study have, therefore, taken account of such variation. Representative specimens for each species, where possible, were taken from the full spectrum of host and environmental conditions available, thus ensuring a robust data set capable of correctly classi-

Table 3. The application of 4 statistical classification methods to morphometric data derived from light microscopy studies of the opisthaptoral hooks and bars (hamulus, marginal hook and ventral bar) of 3 *Gyrodactylus* species

(Abbreviations: *Derj*, *G. derjavini*; *Trut*, *G. truttae*; *Sal*, *G. salaris*; EMR, Estimated Misclassification Rate.)

| | True class | | |
|---|---|---|---|
| Predicted class | *Derj* | *Trut* | *Sal* |
| (A) Nearest Neighbours (NN) | | | |
| *Derj* | 241 | 8 | 0 |
| *Trut* | 10 | 126 | 0 |
| *Sal* | 0 | 0 | 85 |
| EMR = 3·8 % | | | |
| (B) Feed-Forward Neural Network (FFNN) | | | |
| *Derj* | 240 | 11 | 1 |
| *Trut* | 11 | 123 | 0 |
| *Sal* | 0 | 0 | 84 |
| EMR = 5·1 % | | | |
| (C) Projection Pursuit Regression (PPR) | | | |
| *Derj* | 239 | 13 | 1 |
| *Trut* | 12 | 121 | 1 |
| *Sal* | 0 | 0 | 83 |
| EMR = 5·7 % | | | |
| (D) Linear Discriminant Analysis | | | |
| *Derj* | 231 | 43 | 0 |
| *Trut* | 20 | 91 | 0 |
| *Sal* | 0 | 0 | 85 |
| EMR = 13·4 % | | | |

Table 4. The application of the nearest neighbours classification model to data derived from light microscopy studies of individual opisthaptoral structures of 3 *Gyrodactylus* species

(Abbreviations: *Derj*, *G. derjavini*; *Trut*, *G. truttae*; *Sal*, *G. salaris*; EMR, Estimated Misclassification Rate.)

| | True class | | |
|---|---|---|---|
| Predicted class | *Derj* | *Trut* | *Sal* |
| (A) Hamulus subset | | | |
| *Derj* | 239 | 19 | 1 |
| *Trut* | 12 | 115 | 1 |
| *Sal* | 0 | 0 | 83 |
| EMR = 7·0 % | | | |
| (B) Ventral bar subset | | | |
| *Derj* | 209 | 37 | 8 |
| *Trut* | 36 | 93 | 11 |
| *Sal* | 6 | 4 | 66 |
| EMR = 21·7 % | | | |
| (C) Marginal hook subset | | | |
| *Derj* | 225 | 25 | 0 |
| *Trut* | 25 | 109 | 0 |
| *Sal* | 1 | 0 | 85 |
| EMR = 10·9 % | | | |

fying a specimen of *G. salaris* regardless of its origin. The results illustrate the perfect discrimination of *G. salaris* from closely related species infecting salmonids on the basis of a single structure, the marginal hook, measured from SEM-derived images. The perfect discrimination of *G. salaris* was still possible at the level of the light microscope but, to achieve this, more morphometric information was required and thus all 3 opisthaptoral structures were needed. The possibility of a reliable method to provide accurate determinations from data taken with the light microscope when based on data taken from all 3 structures, the hamulus, ventral bar and a marginal hook, presents itself as a candidate system for the rapid, early detection of pathogens.

Of the statistical classification techniques investigated, nearest neighbours consistently provided the lowest EMR values at both the SEM and light level. However, misclassifications resulted when each hook or bar was considered separately using this method on light microscope images. Table 4(C) shows that all *G. salaris* specimens were perfectly identified but 1 specimen of *G. derjavini* was identified as *G. salaris*. It is important to note that, though the lowest EMR value was obtained for the hamulus data subset (3 measured morphometric parameters) (EMR = 7·0%), it was the marginal hook data subset (5 measured morphometric parameters) (EMR = 10·9%) which gave the best classification of *G. salaris*. Therefore, if using a single structure at the light level, the marginal hook subset would appear to be the most useful, since all the submitted *G. salaris* specimens were correctly classified. The classification of specimens using only marginal hook data at the SEM level was based on pooled information from 7 measured morphometric parameters while only 5 parameters were used for the same structure measured using the light microscope. The higher resolution of the SEM images enabled the inclusion and measurement of more morphometric parameters than could be reliably measured using the light microscope. It is highly likely that the EMR values for hooks and bars measured using the light microscope could be reduced further by increasing the number of submitted specimens (i.e. more training data), thereby improving the training capability of the classifier and reducing the probability of misclassification. Alternatively, lowered EMR values may be achieved by increasing the number of morphometric parameters used to describe the morphology of the marginal hook at the light level, where this is possible.

In the application of these statistical classification methods, each specimen has been allocated to that species class which is most probable given the data that describes that particular specimen. It is clearly possible for specimens to be misclassified using such a rule; when additional information regarding the relative seriousness of each type of misclassification is available this can be incorporated into the classification system. No such information has been incorporated in this study. This means that we have implicitly assumed that the costs of the various types of misclassification are equal; thus only the morphometric data are being used to determine the most probable allocation of specimens to classes. We propose, in future work, to determine appropriate estimates of the costs of misclassfications, where possible, and to incorporate them into the classification system. Another issue which has not been implemented to date is that of dealing with new types of specimen which are outside the previous experience of the system. It is possible to use statistical methods for the detection of outliers to identify the occurrence of such specimens (see, for example, Barnett & Lewis (1994)) or for the computation of atypicality indices (see Aitchison & Dunsmore (1975)); such specimens would then be subjected to further scrutiny.

The morphometric characters used for this analysis were selected for the sole purpose of their ability to discriminate *G. salaris* from the other species of *Gyrodactylus* studied (Shinn, 1993; Shinn *et al.* 1996). Thus, the only concern, and the main objective of the study, was the complete and accurate discrimination of *G. salaris*. This, therefore, may explain the origin of most of the misclassifications. It is not surprising that the marginal hooks of *G. salaris* are distinguished more readily in this study than the characteristically shaped marginal hooks of *G. colemanensis* because the input data were selected for this purpose. If, however, the aim of this study was to correctly classify each of the *Gyrodactylus* species, it would possibly require the use of different criteria as input data. At present 7 point to point measurements made on SEM images of the marginal hook or 10 measurements made on 3 structures (the hamulus, ventral bar and a marginal hook), using the light microscope are sufficient to correctly discriminate *G. salaris* from the other species studied here. To achieve the perfect discrimination of all the species of *Gyrodactylus* considered here, input data describing for example the precise shape of the marginal hook sickle, may be required. Current work aimed at producing a package for dissemination will be further strengthened by the expansion of the data training sets, including more specimens for each *Gyrodactylus* species. These specimens will include, where possible, specimens of *G. salaris* from a range of other hosts, including non-salmonids.

The value of statistical classification techniques to solve complex problems in biology have been demonstrated here by the use of the nearest neighbours and the feed-forward neural network methods to provide correct gyrodactylid classification from a single structure at the SEM level. The techniques of linear discriminant analysis and nearest neighbours enabled the discrimination of *G. salaris* using data

taken at the light level. Further, all 3 techniques have been applied in the field of biomedicine as detectors of peripheral vascular disease using arterial pulse waveforms (Allen & Murray, 1996). Nearest neighbours has been used to categorize tissue microcalcifications analysed by X-ray microradiography (Ng, Looi & Bradley, 1996) and to discriminate different types of brain tumours when presented with 3-dimensional magnetic resonance image sequences (Rossmanith *et al*. 1996). Neural networks have been applied in a similar fashion for the early detection of abnormal cancer cells (Moallemi, 1991) and to identify cancer drug candidates and predict their mechanism of action based on databases of drug features (Weinstein *et al*. 1992*a*, *b*; Rouvray, 1993). In biosystematics, nearest neighbours has been used in botany for the discrimination of 13 species belonging to the genus *Pogostemon* (Khanam *et al*. 1994) and in marine biology to identify species of microplankton of the genus *Cymatocylis* (Williams *et al*. 1994). Similarly, neural networks have enabled the identification of poisonous algal species present in phytoplankton samples from co-occurring non-poisonous species (Balfoort *et al*. 1992; Smits *et al*. 1992).

The use and further development of such methodologies will provide novel taxonomic, discriminatory tools for the accurate identification of important pathogens, such as *G. salaris*, and have the potential to be extended to encompass and take account of a wide range of pathogenic and non-pathogenic organisms. The results of this study also suggest that the development of a semi-automated system is feasible whereby the analysis is applied directly to image data, thus avoiding the necessity for manual extraction of measurements.

The authors gratefully acknowledge, for their valuable assistance in the collection of fish samples, Willie Yeomans and the Environment Agency and the numerous fish farms visited during the course of this study. We would also like to thank Dr David Gibson (Parasitic Worms) and the staff of the Electron Microscope Unit (The Natural History Museum, London), Dr James Bron (Institute of Aquaculture, University of Stirling), and the anonymous referees for their constructive comments in the preparation of this manuscript.

## REFERENCES

AITCHISON, J. & DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.

ALDERMAN, D. J. (1996). Geographical spread of bacterial and fungal diseases of crustaceans. *Revue Scientifique et Technique de l'Office International des Epizooties* **15**, 603–632.

ALLEN, J. & MURRAY, A. (1996). Comparison of three arterial pulse waveform classification techniques. *Journal of Medical Engineering and Technology* **20**, 109–114.

BALFOORT, H. W., SNOEK, J., SMITS, J. R. M., BREEDVELD, L. W., HOFSTRAAT, J. W. & RINGELBERG, J. (1992). Automatic identification of algae – neural network analysis of flow cytometric data. *Journal of Plankton Research* **14**, 575–589.

BARNETT, V. & LEWIS, T. (1994). *Outliers in Statistical Data*, 3rd Edn. Wiley, Chichester, New York.

CUNNINGHAM, C. O., MCGILLIVRAY, D. M., MACKENZIE, K. & MELVIN, W. T. (1995*a*). Identification of *Gyrodactylus* (Monogenea) species parasitizing salmonid fish using DNA probes. *Journal of Fish Diseases* **18**, 539–544.

CUNNINGHAM, C. O., MCGILLIVRAY, D. M., MACKENZIE, K. & MELVIN, W. T. (1995*b*). Discrimination between *Gyrodactylus salaris*, *G. derjavini* and *G. truttae* (Platyhelminthes: Monogenea) using restriction fragment length polymorphisms and an oligonucleotide probe within the small subunit ribosomal RNA gene. *Parasitology* **111**, 87–94.

GIBSON, D. I. (1993). *Monobothrium wageneri*: another imported tapeworm established in wild British freshwater fishes? *Journal of Fish Biology* **43**, 281–285.

HARRIS, P. D. (1993). Interactions between reproduction and population biology in gyrodactylid monogeneans – a review. *Bulletin Français de la Pêche et de la Pisciculture* **328**, 47–65.

HAYKIN, S. (1994). *Neural Networks: A Comprehensive Foundation*. MacMillan, New Jersey.

JOHNSEN, B. O. & JENSEN, A. J. (1986). Infestations of Atlantic salmon, *Salmo salar*, by *Gyrodactylus salaris* in Norwegian rivers. *Journal of Fish Biology* **29**, 233–241.

JOHNSEN, B. O. & JENSEN, A. J. (1988). Introduction and establishment of *Gyrodactylus salaris* Malmberg, 1957, on Atlantic salmon, *Salmo salar* L., fry and parr in the River Vefsna, northern Norway. *Journal of Fish Diseases* **11**, 35–45.

JOHNSTON, C., MACKENZIE, K., CUNNINGHAM, C. O., EIRAS, J. C. & BRUNO, D. W. (1996). Occurrence of *Gyrodactylus salaris* Malmberg, 1957, in Portugal. *Bulletin of the European Association of Fish Pathologists* **16**, 89–91.

KAY, J. W., SHINN, A. P. & SOMMERVILLE, C. (1999). Towards an automated system for the identification of notifiable pathogens: using *Gyrodactylus salaris* as an example. *Parasitology Today* **15**, 201–206.

KENNEDY, C. R. (1993). Introductions, spread and colonization of new localities by fish helminth and crustacean parasites in the British Isles: a perspective and appraisal. *Journal of Fish Biology* **43**, 287–301.

KHANAM, M., CULHAM, A., HUSAIN, S. Z. & TAHIR, S. S. (1994). Numerical analysis of the genus *Pogostemon desf* (Lamiaceae). *Pakistan Journal of Botany* **26**, 57–62.

MALMBERG, G. (1957). [On the occurrence of *Gyrodactylus* on Swedish fishes.] *Skrifter utgivna av Södra Sveriges Fiskeriförening Årsskrift 1956*, 19–76. (In Swedish with an English summary.)

MALMBERG, G. (1970). The excretory systems and the marginal hooks as a basis for the systematics of *Gyrodactylus* (Trematoda, Monogenea). *Arkiv för Zoologi, Serie 2*, **23**, 1–235.

MCLACHLAN, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, Chichester, New York.

MO, T. A. (1991*a*). Seasonal variations of opisthaptoral hard parts of *Gyrodactylus salaris* Malmberg, 1957 (Monogenea: Gyrodactylidae) on parr of Atlantic salmon *Salmo salar* L. in the River Batnfjordselva, Norway. *Systematic Parasitology* **19**, 231–240.

MO, T. A. (1991*b*). Variations of opisthaptoral hard parts of *Gyrodactylus salaris* Malmberg, 1957 (Monogenea: Gyrodactylidae) on rainbow trout *Oncorhynchus mykiss* (Walbaum, 1792) in a fish farm, with comments on the spreading of the parasite in south-eastern Norway. *Systematic Parasitology* **20**, 1–9.

MO, T. A. (1991*c*). Variations of opisthaptoral hard parts of *Gyrodactylus salaris* Malmberg, 1957 (Monogenea: Gyrodactylidae) on parr of Atlantic salmon *Salmo salar* L. in laboratory experiments. *Systematic Parasitology* **20**, 11–19.

MO, T. A. (1994). Status of *Gyrodactylus salaris* problems and research in Norway. In *Parasitic Diseases of Fish* (ed. Pike, A. W. & Lewis, J. W.), pp. 43–56. Samara Publishing Ltd, Dyfed.

MOALLEMI, C. (1991). Classifying cells for cancer diagnosis using neural networks. *IEEE Expert – Intelligent Systems and their Applications* **6**, 8–12.

NG, K. H., LOOI, L. M. & BRADLEY, D. A. (1996). Microcalcification clustering parameters in breast disease: A morphometric analysis of radiographs of excision specimens. *British Journal of Radiology* **69**, 326–334.

PLATTEN, M., McLOUGHLIN, M. & SHINN, A. P. (1994). Distribution and identification of gyrodactylid species in fish farms and rivers in Northern Ireland. *Veterinary Record* **135**, 411–412.

ROSSMANITH, C., HANDELS, H., POPPL, S. J., RINAST, E. & WEISS, H. D. (1996). Characterisation and classification of brain tumours in three-dimensional MR image sequences. *Lecture Notes in Computer Science* **1131**, 429–438.

ROUVRAY, D. (1993). Taking a short cut to drug design. *New Scientist* **138**, 35–38.

SHINN, A. P. (1993). The application of new biosytematic techniques in the discrimination of the genus *Gyrodactylus* (Monogenea) on salmonoid fish. Ph.D. thesis, University of Stirling.

SHINN, A. P., GIBSON, D. I. & SOMMERVILLE, C. (1993). An SEM study of the haptoral sclerites of the genus *Gyrodactylus* von Nordmann, 1832 (Monogenea) following extraction by digestion and sonication. *Systematic Parasitology* **25**, 135–144.

SHINN, A. P., SOMMERVILLE, C. & GIBSON, D. I. (1995). Distribution and characterization of species of *Gyrodactylus* Nordmann, 1832 (Monogenea) parasitizing salmonids in the UK, and their discrimination from *G. salaris* Malmberg, 1957. *Journal of Natural History* **29**, 1383–1402.

SHINN, A. P., DES CLERS, S. A., GIBSON, D. I. & SOMMERVILLE, C. (1996). Multivariate analyses of morphometrical features from *Gyrodactylus* spp. (Monogenea) parasitising British salmonids: Light microscope based studies. *Systematic Parasitology* **33**, 115–125.

SMITS, J. R. M., BREEDVELD, L. W., DERKSEN, M. W. J., KATEMAN, G., BALFOORT, H. W., SNOEK, J. & HOFSTRAAT, J. W. (1992). Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. *Analytica Chimica Acta* **258**, 11–25.

S-PLUS 4 (1997). Data Analysis Products Division, Mathsoft, Seattle.

VENABLES, W. N. & RIPLEY, B. D. (1997). *Modern Applied Statistics with S-Plus*, 2nd Edn. Springer-Verlag, Berlin.

WEINSTEIN, J. N., MYERS, T., BUOLAMWINI, J., RAGHAVEN, K., VANOSOL, W., LICHT, J., VISWANADHAN, V. N., KOHN, K. W., RUBISTEIN, L. V., KOUTSOUKOS, A. D., MONKS, A., SCUDIERO, D. A., ANDERSON, N. L., ZAHAREVITZ, D., CHABNER, B. A., GREVER, M. R. & PAULL, K. D. (1992*a*). Predictive statistics and artificial intelligence in the US National Cancer Institutes drug discovery program for cancer and AIDS. *Stem Cells* **12**, 13–22.

WEINSTEIN, J. N., KOHN, K. W., GREVER, M. R., VISWANADHAN, V. N., RUBINSTEIN, L. V., MONKS, A., SCUDIERO, D. A., WELCH, L., KOUTSOUKOS, A. D., CHIAUSA, A., YANG, J., SCHIFF, R., LICHT, J., WITTES, R. E. & PAULL, K. D. (1992*b*). Neural computing in cancer drug development – predicting mechanism of action. *Science* **258**, 447–451.

WILLIAMS, R., McCALL, H., PIERCE, R. W. & TURNER, J. T. (1994). Speciation of the tintinnid genus *Cymatocylis* by morphometric analysis of the Loricae. *Marine Ecology Progress Series* **107**, 263–272.

YEOMANS, W. E., CHUBB, J. C. & SWEETING, R. A. (1997). *Khawia sinensis* (Cestoda: Caryophyllidea) – an indicator of legislative failure to protect freshwater habitats in the British Isles? *Journal of Fish Biology* **51**, 880–885.