Original Article

Discriminant analysis: A supportive tool for monogenoidean taxonomy

Nirupama Agrawal[1,*], Girdhar G. Agarwal[2], Priyanka Tripathi[1,*], Rashmi Pant[2]

[1] *Department of Zoology, University of Lucknow, Lucknow, India;*
[2] *Department of Statistics, University of Lucknow, Lucknow, India.*

Summary **Data on the measurements of haptoral parts of 47 parasite species of *Bychowskyella* (*1*), *Cornudiscoides* (*2*) and *Thaparocleidus* (*3*), found on Indian freshwater catfish, were gathered from previously reported studies. Based on six morphometric variables concerning haptors, these species were subjected to discriminant analysis in order to more accurately distinguish their generic placement. This paper describes a comparative study of several classification models. Using the original variables in the study, a simple linear discriminant analysis model was constructed and models using principal components (PC) for discrimination have also been explored. The effectiveness of these models is measured in terms of percentage of correct classification. Quadratic discriminant models using original variables and their principal components provided the highest (95%) correct classification for *Bychowskyella* (*1*). Species of *Thaparocleidus* (*3*) were correctly classified using dorsal and ventral hard-part measurements (100%) as well as hook measurements. The highest percentage (75%) of correct classification for *Cornudiscoides* (*2*) was achieved through a quadratic discriminant model using hook measurements.**

Keywords: Discriminant analysis, Principal component analysis, Monogenoidea, Taxonomy, Hard part

1. Introduction

Identification of parasites using only morphological techniques has long been debated by molecular biologists, statisticians, and related researchers. Analysis (characterization) of monogenoideans has often remained an arduous task for the researchers in the field. Their grouping into different taxa has long been questioned. Misidentification and misinterpretation often creates problems with their classification and thus leads to incorrect phylogenies. Many of the Indian species were originally misplaced in incorrect genera due to poor understanding of the morphological boundaries of dactylogyrid genera during early studies in India (*4*). This confusion can be avoided if assistance is sought from other fields in order to reach the correct classification. Distinguishing parasites will achieve greater accuracy if appropriate tools are provided by statistics. This will help to clarify the taxonomic status of monogenoidean parasites. In the present work, statistical analysis has been used to supplement the taxonomical work carried out to date regarding the categorization of the species of three genera *Bychowskyella* (*1*), *Cornudiscoides* (*2*), and *Thaparocleidus* (*3*) belonging to the family dactylogyridae in the class monogenoidea. This analysis of morphometric variation is used to assess the congruence of genera and species initially distinguished by morphology, where species are recognized based on morphometric variation of haptoral sclerites. A review of monogenoidean parasites of siluriformes fishes from the Old World (*5*) indicated that 17 species of *Bychowskyella* (*1*), 19 species of *Thaparocleidus* (*3*), and 11 species of *Cornudiscoides* (*2*) from India are considered valid. The validity of certain species of these genera is still disputed. To overcome this problem, this work uses discriminant analysis, a numerical taxonomy technique that divides data into groups so that subjects within a group are similar to one another and differ from subjects in other

**Correspondence to:* Dr. Nirupama Agrawal, Dr. Priyanka Tripathi, Department of Zoology, University of Lucknow, Lucknow-226007, Uttar Pradesh, India;
e-mail: dr_neeru_1954@Yahoo.co.in
e-mail: Priyanka_tripathi1980@yahoo.com

groups. Relationships among measurement variables with respect to the grouping variable can be expressed by their mean values and variance-covariance matrices. Discriminant analysis has been used in monogenoidean work by previous researchers (*6*) who examined hook, ventral bar, and anchor features of only four *Gyrodactylus* species: *G. gondae*, *G. flavescensis*, *G. arcuatus*, and *G. arcuatoides*. Approximately 8-23 specimens per species were measured. This is a rather low number for proper statistical analyses. Other researchers (*7-9*) discriminated between only 2 closely related and morphologically similar species of *Gyrodactylus*; *G. salaris* and *G. thymalli*, using the statistical classification methodologies of linear discriminant analysis (LDA) and k-nearest neighbors (KNN). To overcome the shortcomings of earlier research, an attempt has been made here to select as many variables as possible for accurate analysis. The goal of this paper is to investigate the role of discriminant analysis in conjunction with principal component analysis in order to validate the species of three monogenoidean genera: *Cornudiscoides* (*2*), *Thaparocleidus* (*3*), and *Bychowskyella* (*1*).

2. Materials and Methods

Data of the hard parts of 47 parasite species belonging to three genera *Cornudiscoides* (*2*), *Thaparocleidus* (*3*), and *Bychowskyella* (*1*) were compiled from published studies and subjected to discriminant analysis. Several published and ongoing studies concerning monogenoidean communities were reviewed. Morphometrical distances of haptor sclerite parts were used. The terminology followed here is that of Gusev (1976) (*10*). This study provided a set of quantitative observations on several species of parasites. This was done to classify organisms into taxonomic categories based on their quantitative measurements.

Discriminant analysis is a statistical technique in which the quantitative measurement of cases is used to create a model that explains the classification of the cases into different groups. This model can further be used to assign additional observations to the correct group. Such models can be fit in a variety of ways depending on the covariance or correlation structure of the variables. In the simplest case, when the covariance matrices are equal for each group, the linear discriminant model (function) is used. In the most general case, when covariance matrices are unequal for the groups, the discriminant function is quadratic (*i.e.* quadratic discriminant analysis or QDA) (*11*).

A classical linear discriminant analysis (LDA) of a three-class outcome with two or more feature variables results in two linear discriminants that are linear combinations of the features. A scatter plot of these two variates, with the data points marked by class, shows the effectiveness of the discrimination between classes. Sometimes when the number of variables is large or there is multicollinearity (*12*) among the feature variables, the analysis can be simplified by considering a smaller number of linear combinations of the original variables (*13*). Principal component analysis (PCA) is a technique that finds such linear combinations (called principal components). Usually the first two or three principal components explain most of the variation in the original data. This paper uses a classification method based on PCA and DA. The method consists of two steps: first, the original vector space is projected to a subspace *via* PCA, and then DA is used to obtain a best classifier. The basic idea of combining PCA and DA is to improve the generalization capability of DA when only few samples per class are available. S-PLUS and SPSS software were used for statistical processing in this study.

3. Results

3.1. *Statistical analysis*

The variables in this study were (measurements in microns):

1. Dorsal Anchor inner length (DALI)
2. Dorsal Anchor outer length (DAO)
3. Dorsal Bar length (DBL)
4. Ventral Bar length (VBL)
5. Ventral Anchor length inner (VALI).
6. Hook measurements (H1, H2, H3, H4, H5, H6, H7)

The classes (groups) are described by the variable GENUS.

A univariate analysis of variance (Table 1) indicated a significant difference between the three group means with respect to dorsal bar length ($p = 0.007$), ventral bar length ($p = 0.006$), and ventral anchor inner length ($p < 0.001$). Although the differences in group means for DALI were not significant ($p = 0.301$), they might serve as an ancillary variable and provide, along with significant variables, additional information for classification purposes.

The relationships between the variables used for classification should be explored.

The matrix plot in Figure 1 shows the direction of relationships between the original variables. The last row and column indicate the differences between the three genera. As is apparent, the variables for

Table 1. Mean and standard deviation (Mean; SD) of variables with respect to each genus

	Bychowskyella ($n = 20$) Mean; SD	*Cornudiscoides* ($n = 12$) Mean; SD	*Thaparocleidus* ($n = 15$) Mean; SD	*p*-value
DALI	70.63; 23	97.82; 174.8	43.8; 13.21	0.301
DBL	50.98; 25.57	31.75; 5.17	34.3; 11.03	0.007
VALI	40.29; 17.33	25.25; 9.55	21.63; 6.35	<0.0001
VBL	63.1; 43.76	36.92; 6.44	31.27; 10.05	0.006

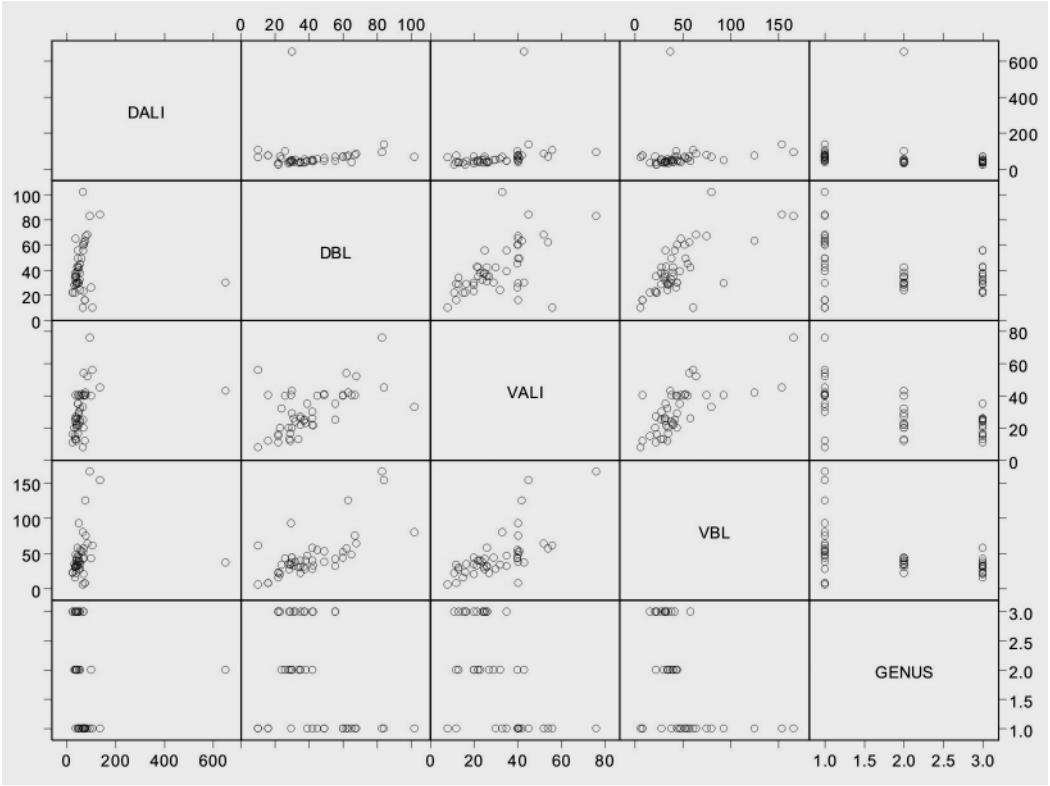

Figure 1. Matrix plot indicating strength and direction of relationship between variables.

Bychowskyella (*1*) have a markedly different location and scatter pattern while *Cornudiscoides* (*2*) and *Thaparocleidus* (*3*) have a similar pattern.

An important assumption in linear LDA is that the within groups covariance matrices should be equal for each group. The differences in covariance matrices for each of the three groups were found to be highly significant ($p < 0.05$). Hence, the data suggests a heteroscedastic covariance structure.

3.2. *Construction of discriminant models*

Next, the classification obtained by linear and quadratic classifiers was compared using original variables. The best overall classification is obtained by using a quadratic discriminant function with original variables as independents (Table 2). An overall correct classification of 80.9% is obtained. While the *Bychowskyella* (*1*) and *Thaparocleidus* (*3*) species are correctly classified, the classification of *Cornudiscoides* (*2*) is merely 33.3%.

In order to determine the maximum separation between *Cornudiscoides* (*2*) and *Thaparocleidus* (*3*), the two were studied separately. Since dorsal anchor outer length is an additional feature available for the two genera, it was also incorporated in the classification model. The matrix plot in Figure 2 shows the direction of relationships between the previous four variables and the additional variable (dorsal anchor outer length) for the two genera.

Table 2. Comparison of various linear and quadratic discriminant models

Genus	Correct classification (%) Model	
	LDA	QDA
Bychowskyella	90	95
Cornudiscoides	8.4	33.3
Thaparocleidus	94	100
Overall	70.21	80.9

LDA, linear discriminant analysis; QDA, quadratic discriminant analysis.

The best model for the discrimination of species in the two groups, based on dorsal and ventral hard parts, is the linear discriminant model with the separate principal components of the dorsal and ventral hard parts. The first principal component of the dorsal anchor outer length and dorsal bar length accounts for 80% of variability in the data and the first principal component of the ventral anchor inner length and ventral bar length also accounts for 80% of variability in the data. Discriminant analysis using these two components as new predictors yielded the following classification, as shown in Table 3.

The apparent error rate (APER) for the above classification was 29.6%, which is quite reasonable. The linear discriminant function is shown in Figure 3.

In order to find an even better way to classify the two genera *Cornudiscoides* (*2*) and *Thaparocleidus* (*3*), measurements of the seven pairs of hooks in each species were also used. The best classification was

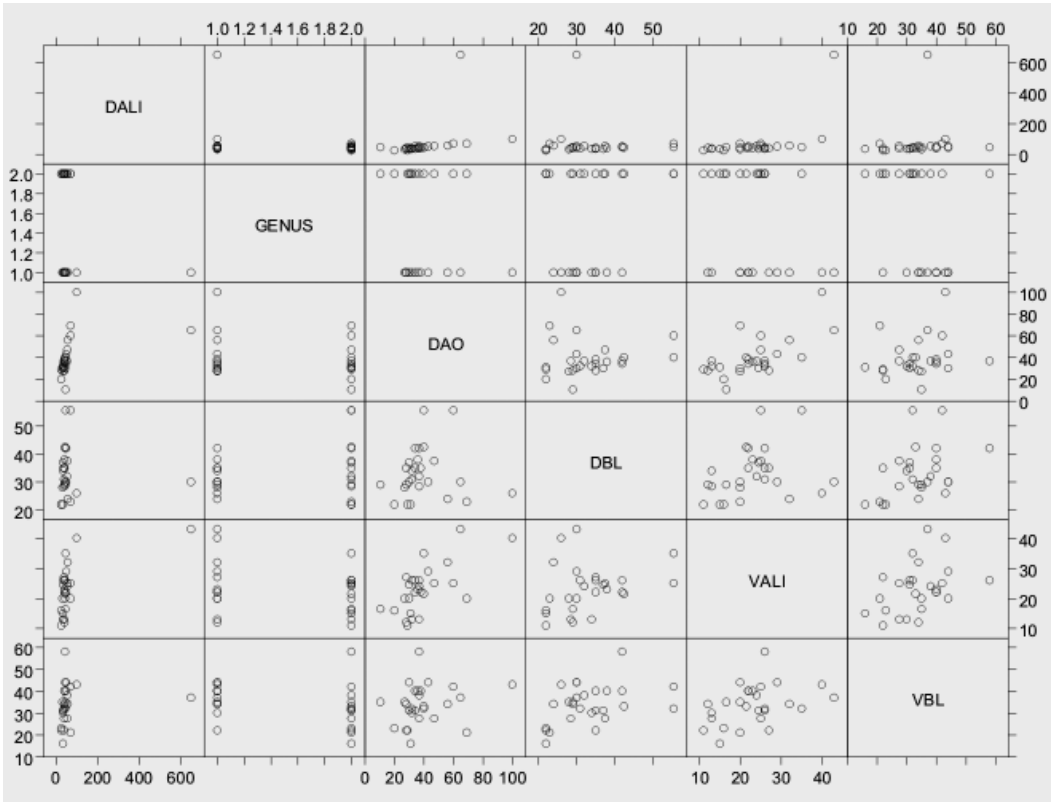

Figure 2. Relationships between variables for genus *Cornudiscoides* (*2*) and *Thaparocleidus* (*3*).

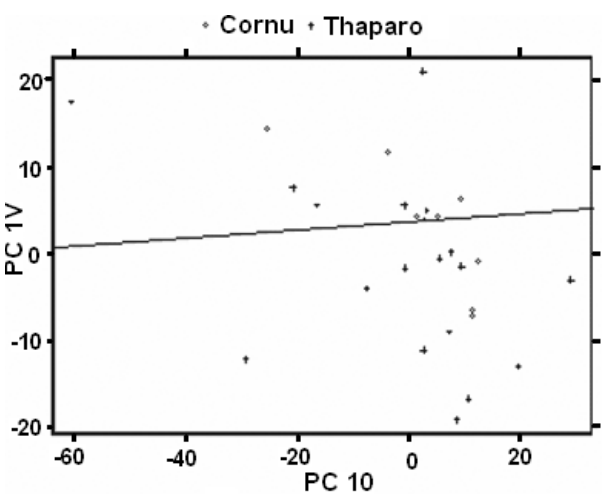

Figure 3. Linear discriminant function using two principal components.

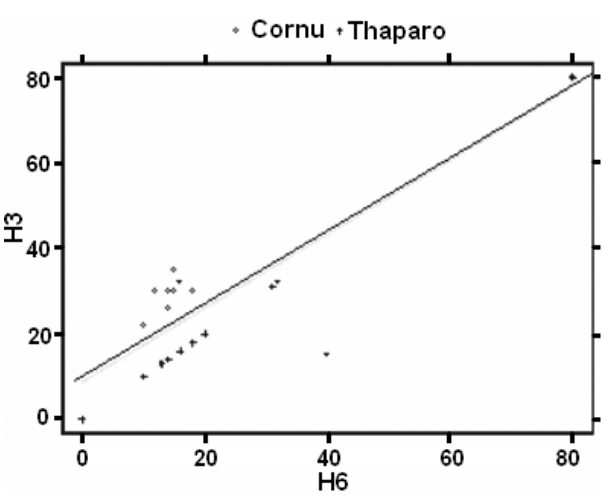

Figure 4. Quadratic discriminant analysis using hooks data. H3 is the third hook and H6 is the sixth hook.

Table 3. Summary of LDA using principal components of dorsal and ventral hard parts

		Predicted genus		Total number of parasites
		Cornudiscoides	*Thaparocleidus*	
Actual genus	*Cornudiscoides*	8	4	12
	Thaparocleidus	4	11	15

Table 4. Summary of QDA using hook measurements

		Predicted genus		Total number of parasites
		Cornudiscoides	*Thaparocleidus*	
Actual genus	*Cornudiscoides*	9	3	12
	Thaparocleidus	0	15	15

found to be a quadratic discriminant function depending on hook variables (H3 and H6) with overall correct classification of 88.9% (Table 4).

The highest percentage (75%) of correct classification for *Cornudiscoides* (*2*) and 100% correct classification for *Thaparocleidus* (*3*) genera were achieved through the QDA model. The quadratic discriminant function is shown in Figure 4.

4. Discussion

The order Siluriformes, or catfishes, includes 13 genera and about 100 species of freshwater fishes and has a wide geographical distribution, being found in Africa, Syria, and southern and western Asia (Philippines to Java) (*14*). Parasites classified based on statistical technique are found on catfishes that belong to this order. Discrimination among these parasites was required, necessitating their critical reexamination. The current analysis identified significant differences ($p < 0.05$) between species within each genus for six measured haptoral sclerite characteristics. Quadratic discriminant models using original variables and their principal components provided the highest (95%) correct classification for the genus *Bychowskyella* (*1*). The single species that remains incorrectly classified in *Bychowskyella* (*1*) is *Bychowskyella tripathii* (*15*). The species of *Thaparocleidus* (*3*) were correctly classified using dorsal and ventral hard part measurements (100%) as well as hook measurements. The highest percentage (75%) of correct classification for *Cornudiscoides* (*2*) was achieved through a QDA model using hook measurements. The three species that are incorrectly classified in *Cornudiscoides* (*2*) are *Cornudiscoides heterotylus* (*2*), *Cornudiscoides megalorcis* (*2*), and *Cornudiscoides geminus* (*10*). These species need to be re-examined. Any method of classification should thus take into account the heteroscedasticy in variables for the different groups. Some hard parts were also observed to serve as a more effective method of classification (*e.g.* hooks). Therefore, such methods should identify morphometric characteristics that best identify species belonging to the three genera.

Statistical implications of the measurements of these parasites are that taxonomists, who consider *Bychowskyella* (*1*), *Cornudiscoides* (*2*), and *Thaparocleidus* (*3*) to be three different and well-established genera, are correct in their claim. Thus, their correct interpretation offers potential for the correct mapping of parasite phylogeny. The blurred distinctiveness of these taxa was restored by subjecting them to discriminant analysis. The fact that the three discriminated genera were quite different from each other and not statistically similar posed no problem whatsoever.

Acknowledgements

The work was supported by a grant (No. CST/AAS/D-03) from the State Council of Science & Technology (UP).

References

1. Achmerow AKH. New species of monogeneans from fishes of Amur River. Parazitolog Sbornik 1952; 4:181-212.
2. Kulkarni. Studies on the monogenetic trematodes of fishes found in Hyderabad, Andra Pradesh (India). Part I Rivista di parassitologia 1969; 30:73-90.
3. Jain SL. Monogenea of Indian freshwater fishes. II. *Thaparocleidus wallagonius* n.g., n. sp. (Subfamily: Tetraonchinae) from the gills of *Wallagonia attu* (Bloch), from Lucknow. Indian J Helminthol 1952; 4:43-48.
4. Kritsky DC, Pandey KC, Agrawal N, Abdullah MAS. Monogenoids from the gills of spiny eels (Teleostei: Mastacembelidae) in India and Iraq, proposal of *Mastacembelocleidus* gen.n., and status of the Indian species of *Actinocleidus*, *Urocleidus* and *Haplocleidus* (Monogenoidea: Dactylogyridae). Fol Parasitol 2004; 51:291-298.
5. Lim LHS, Timofeeva TA, Gibson DI. Dactylogyridean monogeneans of the siluriform fishes of the old world. Syst Parasitol 2001; 50:159-197.
6. Huyse T, Malmberg G, Filip AM, Volckaert FAM. Four new species of *Gyrodactylous* von Nordmann, 1832 (Monogenea, Gyrodactylidae) on gobiid fishes: combined DNA and morphological analyses. Syst Parasitol 2004; 59:103-120.
7. Corlis D. Four new genera of Monogenea (Dactylogyridae) from the gills of Australian atheriniform freshwater fishes: Memoir Queensl Mus 2004; 49:537-571.
8. Strona G, Stefani F, Benzoni F, Harhash KA Eman S, Galli P. Morphometric discriminant analysis for the classification of *Diplectanum* (Monogenea: Monopisthocotylea), parasites of *Sphyraena flavicauda*. Parasitol 2005; 47:237-240.
9. McHugh ES, Shinn AP, Kay JW. Discrimination of the notifiable pathogen *Gyrodactylus salaris* from G. *thymalli* (Monogenea) using statistical classifiers applied to morphometric data 2000; 121 (Pt 3):315-323.
10. Gusev AV. Freshwater Indian Monogenoidea. Principles of systematics, analysis of world faunas and their evolution. Indian J Helminthol 1976; 25 & 26:1-241.
11. Huberty CH. Applied Discriminant Analysis. New York, John Wiley & Sons 1994.
12. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis Academic Press, London, 1979.
13. Flury B. Common Principal Components in k groups. J American Statist Assoc 1984; 79:892-898.
14. Froese R, Pauly D. Fishbase, World Wide Web electronic publication. *www.Fishbase.org*, version (01/ 2007).
15. Kumar R, Agarwal GP. On a new monogenetic trematode, *Bychowskyella tripathii* n.sp. from the gills of a freshwater fish Wallago attu (Bl. & Sch.). Jpn J Parasitol 1981; 30:1-8.

(*Received May 30, 2008; Accepted June 27, 2008*)