Identification of Monogenea made easier: a new statistical procedure for an automatic selection of diagnostic linear measurements in closely related species

GIOVANNI STRONA1, SIMONE MONTANO1, DAVIDE SEVESO1, PAOLO GALLI1 and SIMONE FATTORINI1,2

Abstract

We introduce a new statistical method to select which morphological characters are most useful to identify monogenean species. The method estimates the average size overlap of candidate diagnostic structures among a set of species to individuate those that mostly differ between the species. To demonstrate our approach, we report a comprehensive analysis conducted on two of the most species-rich monogenean genera: Dactylogyra Diesing, 1850 and Gyrodactylus von Nordmann, 1832. We demonstrate that, in contrast to common taxonomic practice, very few but highly diagnostic measurements are necessary to correctly identify a specimen. In particular, we found that most of Dactylogyra and Gyrodactylus species can be classified on the basis of the width of the supplementary connecting bar and of the length of the hook sickle, respectively.

Key words: Evolutionary radiation – morphometry – regression tree – sclerites

Introduction

Some 4000 monogenean species are currently known, and probably some 25 000 species exist (Whittington 1998). Monogenea are a morphologically very diversified group, being the only flatworm clade that underwent a wide adaptive radiation (Brooks and McLennan 1993), with the evolution of a great variety of structural designs in the attachment organs (Kearn 1994), which are frequently used to identify species. In particular, the posterior attachment organ, called opisthaptor, is characterized by the presence of various sclerotized structures (anchors, bars, hooks, etc.) whose number, morphology and morphometry are extremely variable between monogenean species (Boeger and Kritsky 1993) and are used as diagnostic characters in taxonomy (Vignon 2011a,b). However, in some groups, such as Gyrodactylidae, species identification based on morphometry of sclerites is difficult due to the small size and low variability in these characters for closely related species (Shinn et al. 2010).

The widespread use of sclerites is also related to the fact that, thanks to their chemical composition, they are the least perishable monogenean structures, allowing species identification even for specimens that are not optimally preserved. This is important, because the best way to prepare monogenean slides for long-term preservation is debated, and incorrect preparation may produce serious morphological artefacts (Wong et al. 2006; Strona et al. 2009; Kosková et al. 2010; Justine et al. 2012).

However, not all sclerotized structures can be clearly visible on prepared materials, because of different reasons, such as a suboptimal orientation of specimen on slide, or the chemical properties of the media used for preservation (Galli et al. 2007). For example, one of the most common solutions used to fix monogenean parasites is ammonium pycrate-glycerine (Malmberg 1956), which is, however, known to continue clearing the specimens long after fixation, eventually turning some of the internal structures from hardly detectable to almost completely invisible (Kritsky et al. 1978). This makes it difficult, or even impossible, to identify a specimen on the basis of the presence–absence or of the precise number of particular sclerites.

The use of a certain number of measurements taken on visible sclerites (Gussev 1976) makes it possible to overcome this issue and to correctly identify a specimen on the basis of structures that are diagnostic and generally less perishable in museum materials. But even this approach is not always completely safe, because sclerites could have been damaged during the preparation of slides and because they are unlikely to be oriented on the same plane of the slide, which may lead to mistakes in measurements. This has been clearly highlighted by the use of confocal microscopy, which has demonstrated how these two issues may, respectively, lead to over- and under-estimation of the true size of sclerites (Galli et al. 2006, 2007).

Gussev (1976) has indicated a complex system of measurements to correctly identify monogenean species. However, several studies demonstrated that many morphometric variables are highly intercorrelated, which diminishes their value to discriminate species (Shinn et al. 1996; Du Preez and Maritz 2006; Vignon 2011a,b). This problem is particularly important in closely related species (Strona et al. 2005). Because Monogenea include several genera with tens of species (Whittington 1998), identifying highly discriminant morphometric variables is particularly compelling.

To cope with colinearity among morphological measurements, Vignon (2011a,b) proposed a combination of dimensions’ reduction using partial least squares with traditional classification methods for morphological data. This approach, however, implies the use of a great number of measurements, most of which might eventually be of little value. Here, we describe a new statistical protocol to identify the smallest set of measurements necessary to confidently assign an unidentified specimen to a particular monogenean species among a set of closely related ones. Our approach differs substantially from that of Vignon (2011a,b), because it is based on size overlap and thus requires only the range (maximum – minimum length) of the considered measurements for each species under study. For purpose of demonstration, we report the results of an extensive application of our method to the genera Dactylogyra Diesing, 1850 and Gyrodactylus von Nordmann, 1832, which are the most species-rich genera of Monogenea (Gibson et al. 1996; Harris et al. 2004).
Material and Methods

Assumptions and basic reasoning

Let us consider a set of parasite species provided with a particular sclerotized structure with a size that varies both within and between species. We can imagine to make a random measurement of the considered structure with the narrowest overlap between the considered species; that is, that the difference between maximum and minimum values, for that particular structure. We can imagine to generate a random value falling within the size range of the structure for a particular parasite species (hereafter ‘P’). Because this value is randomly generated, that is, it is extracted at random from a uniform distribution, we are not making any assumption about the statistical distribution of the actual measurements. Then, we can imagine to take into account the size ranges of the same structure in the whole set of parasite species (including P) and to compare them to the simulated value. The simulated value will fall, of course, within the size range of P. However, we cannot exclude that the value could also fall within the size ranges of the same structure in other parasite species, depending on the extent of size range overlap among species. The larger the number of species with a range including the simulated value, the lower the discriminant power of the selected structure with respect to P. In other words, if we have a specimen belonging to the species P, the size of the above-mentioned structure would not provide a valuable information to correctly assign that specimen to P, as there is a good chance that it could also fall within the size ranges known for other parasite species.

We can now imagine to replicate the same experiment by generating a random measurement of the considered structure for any other parasite species of the set and then check how many species have a size range capable to include each simulated value. By doing this, we would obtain a general estimate of size range overlap for the considered structure. To avoid potential biases due to the extraction of values too close to the extremes of a given range, such estimate could be improved by extracting at random many simulated values from the known range of each considered species.

Then, we can imagine to consider another structure and thus reiterate the above-described procedure using at once two measurements; that is, as the morphometric measurement that performed best about the statistical distribution of the actual measurements. Then, we can imagine to consider another structure and thus reiterate the above-described procedure using at once two measurements; that is, as the morphometric measurement that performed best about the statistical distribution of the actual measurements. Then, we can consider the average overlap between species (as intended above) is likely to be reduced by the addition of the second measurement. More in general, the overlap is likely to be progressively reduced by any subsequent addition of a new measurement, as we can reasonably expect that this would not have the same size range distribution in all the considered species. However, these additions would have different effects on the overlap, depending on how much the considered structures differ between species. Thus, testing different combinations of measurements would allow the identification of the minimum set of morphometric characters that mostly reduce the between-species overlap, that is, it would allow the identification of those characters with the highest discriminant power.

Explanation of the procedure

The proposed method only requires a minimum amount of information for each species, that is, the size range (minimum and maximum value) of the diagnostic sclerotized structures that have to be tested for discriminating power. The procedure is based on the following steps:

1. For each considered species, a set of simulated specimens is created, by assigning to each of them a set of random measurements falling within the known range sizes (minimum-maximum) of the sclerotized structures under examination.
2. A subset of all possible combinations of the generated values of measurement for the examined structure is selected.
3. Each simulated specimen is then compared to each parasite species under examination (i.e. both the species whose size ranges were used to create the simulated specimen, and all related species).
4. An ‘individual score’ (sc) is computed for each simulated specimen, as the total number of times that all measurements belonging to the subset selected at point 2) fall within the corresponding known ranges of a parasite species.
5. An ‘overall score’ (SC) is computed as the total number of simulated specimens quoted by the sum of all sc values. SC is the reciprocal of the average number of species to which a simulated specimen might be assigned on the basis of the selected set of measurements (see point 2); it can therefore vary between 1 (when each simulated specimen is assigned to only its ‘correct’ species) and 1/N, where N is the total number of considered species (when the measurements of each simulated specimen fall within the corresponding ranges of any considered known species);
6. Points 2 to 5 are reiterated for each possible combination of measurements, and a SC value is therefore associated with each possible combination of measurements;
7. A regression tree is constructed using as independent variables the respective presence–absence value (1-0) of each measurement in any possible simulated combination, and as dependent variable the corresponding SC value. Regression trees recursively partition response variables into subsets based on their relationship with a certain number (one to many) predictor variables. Thus, regression trees provide a powerful yet simple technique to uncover dependencies among predictor variables when data interact in complex, nonlinear ways, making it difficult to create a global model (De’ath and Fabricius 2000).

Case study

To show the functioning and the potential of our approach, we performed two sets of analyses using 149 species belonging to Dactylogyrus and 417 species belonging to Gyrodactylus, with the aim of identifying which measurements (among the 11 reported in Fig. 1A) are the most useful to discriminate between Dactylogyrus species and which measurements (among the 10 reported in Fig. 1B) are the most useful to discriminate between Gyrodactylus species. Size ranges for the considered sclerotized structures were obtained from the study by Pugachev et al. (2009) and are reported in Table S1. Analyses were conducted using the following procedure.

For each species, we created 1000 simulated specimens, that is, we created 1000 sets of random values for 11 morphometric characters and 1000 sets of random values for 10 morphometric characters falling within the corresponding known size ranges for Dactylogyrus and Gyrodactylus, respectively. Then, for any possible combination of measurements (2047 for Dactylogyrus and 1024 for Gyrodactylus), we compared the range sizes of each simulated specimen with the corresponding known range size of each considered parasite species, to compute the respective SC values. Finally, for both Dactylogyrus and Gyrodactylus, we generated a regression tree using the SC scores as dependent variable and the corresponding presence–absence of measurements as independent variables.

Creation of simulated specimens and the computation of SC values were carried out using self-developed scripts written in Python programming language (van Rossum and de Boer 1991), whereas the regression tree was generated using the R package ‘rpart’ (Therneau et al. 2013). All scripts and detailed instructions necessary to replicate the above-described analyses are provided as Supporting Information in Table S2.

Results

Regression trees describing the relative discriminant power of each measurement shown in Fig. 1A,B are presented in Fig. 2A, B respectively. Width of the supplementary connecting bar was identified as the morphometric measurement that performed best to discriminate between Dactylogyrus species, that is, the sclerotized structure with the narrowest overlap between the considered species. Another important measurement for the identification of Dactylogyrus species was the overall length of the copulatory organ. Using only these two measurements, it was possible to assign, on average, each simulated specimen to less than 1.2 Dactylogyrus species (including the ‘correct’ one, that is, that used to generate the random values). For Gyrodactylus, the morphometric character that performed best to discriminate between
the considered species was the hook sickle length. Other important characters were the ventral bar length, the membrane length and the dorsal bar length. Using only these four measurements, it was possible to assign, on average, each simulated specimen to less than 1.5 Gyrodactylus species (including the 'correct' one).

The graphs in Fig. 3A,B report the relationships between the number of measurements present in each combination and the average number of species each simulated specimen is attributed to (including the correct one) on the basis of the selected set of characters for the two genera. This relationship is asymptotic; thus, after a certain number of characters have been included, addition of further characters does not reduce the average number of species to which a simulated specimen is attributed (i.e. it does not improve the SC).

Discussion

Because of strong morphological similarities, morphometric identification of closely related species requires the use of characters with high diagnostic power, that is, characters that have minimum overlap among species. However, the choice of such 'diagnostic' characters is often based only on taxonomists' subjective expertise and lacks robust statistical investigation (see Vignon 2011a,b for a detailed discussion on Monogenea). The method described here provides a possible general solution to this issue. Although we tested it on Monogenea species, our procedure may be applied to any other group in which species are identified according to similar morphometric criteria, for example to sibling arthropod species that are identified by subtle differences in the measurements of sclerotized parts of male genitalia (Wakeham-Dawson et al. 2004; Dapporto 2007). One of the main advantages of our procedure in respect to traditional multivariate discriminant methods (Strona et al. 2005; Vignon 2011a, b) is that it requires minimal knowledge of the morphology of the species under examination and of their respective diagnostic structures, namely the size range (minimum and maximum values) of the morphological traits used. This type of information is usually available from taxonomical papers or can be easily obtained from voucher specimens.
The discriminant power of each considered structure depends on the extent of overlap of range size among the species under examination. We cannot exclude that, in some cases, all the measurements traditionally considered as diagnostic could have no overlap among species, which would make their identification obvious and our technique pointless; however, this is a very improbable situation (especially in closely related species, which are the main target of our procedure) because functional and evolutionary constraints are likely to promote a certain degree of similarity between species (Poisot et al. 2011), that is, it is more probable that some structures will show substantial overlap among species, while others do not, thus making the identification of the most valuable morphometric characters important. This is consistent with what we observed in Dactylogyrus and Gyrodactylus species.

In Dactylogyrus and Gyrodactylus, no combination of measurements (including those comprising all measurements) led to a SC value equal to one. This means that all considered measurements for both genera have some overlap in range size among species. However, in both genera, several different combinations of measurements led to SC values >0.5, that is, they made it possible to assign each simulated species to less than two known species (including the correct one, that is, the species whose size ranges were used to generate the simulated specimen) (Fig. 3A, B). In addition, SC values >0.5 were achieved with a small number of variables, with a very slight improvement deriving from the addition of further variables. These results indicate that the common procedure of taking several measurements to identify a monogenean species (Gussev 1976) is not only time-consuming, but also does not ensure that the most valuable measurements are taken. Use of empirical asymptotic curves as those produced would be otherwise hardly detectable. We used regression trees because they are characterized by a high detection power and are very simple in both their implementation and interpretation. However, it should be highlighted that they are not based on a probabilistic model, and thus, there are no probability levels or confidence intervals associated with their outcome (De’ath and Fabricius 2000). Yet, we do not exclude that other techniques of multivariate analyses could be used to model the relationship between character combinations and SC scores, to broaden the fields of application of our approach. However, in our purposes, the described technique is aimed at providing taxonomists with a practical tool to properly focus their research efforts and should therefore be handled with care when used for predictions. Nonetheless, we hope that the increasing availability of online sources for parasitological data (see, for example, Strona and Lafferty 2012; Strona et al. 2013) would provide a framework for better testing our approach, by monitoring, for example, how it responds to the addition of information coming from descriptions of new species. Moreover, our method offers insights into general issues related to the functions of the investigated structures, by promoting further research on the causes of size variability. In this sense, it has broad potential in the field of evolutionary ecology, as it would make it possible to perform robust studies on evolutionary radiation at a very fine scale, by focusing on morphometrical differences that would be otherwise hardly detectable.

Acknowledgements

This work has been supported by the Department of Biotechnology and Biosciences of University of Milano-Bicocca (Grant Number: 2-18-599000-6; U.A. 001; CUP: H47E11000010002). We are grateful to Pavel Gerasev for his useful comments on the manuscript.

Author contributions

G.S. conceived the idea, G.S. and S.F. designed most of the study; P.G, S.M. and D.S. designed part of the study. P.G. provided data. G.S. developed the scripts and performed most of the analyses. S.F. performed some of the analyses. G.S. and S.F. wrote the article.

References

Identification of Monogenea made easier


Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Size ranges of the sclerotized structures considered in the analyses, as reported in the study by Pugachev et al. (2009).

Table S2. Python and R code necessary to replicate the method described in the paper.

© 2013 Blackwell Verlag GmbH