00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 35: 481–486, 2012
doi: 10.1111/j.1600-0587.2012.07439.x

Subject Editor: Thiago Rangel. Accepted 13 January 2012

How to catch a parasite: Parasite Niche Modeler (PaNic) meets Fishbase

Giovanni Strona and Kevin D. Lafferty

G. Strona (giovanni.strona@unimib.it), Dept of Biotechnology and Biosciences, Univ. of Milano-Bicocca, IT-20126 Milano, Italy. – K. D. Lafferty, U.S. Geological Survey, Western Ecological Research Center, Marine Science Inst., Univ. of California, Santa Barbara, CA 93106, USA.

Parasite Niche Modeler (PaNic) is a free online software tool that suggests potential hosts for fish parasites. For a particular parasite species from the major helminth groups (Acanthocephala, Cestoda, Monogenea, Nematoda, Trematoda), PaNic takes data from known hosts (maximum body length, growth rate, life span, age at first maturity, trophic level, phylogeny, and biogeography) and hypothesizes similar fish species that might serve as hosts to that parasite. Users can give varying weights to host attributes and create custom models. In addition to suggesting plausible hosts (with varying degrees of confidence), the models indicate known host species that appear to be outliers in comparison to other known hosts. These unique features make PaNic an innovative tool for addressing both theoretical and applied questions in fish parasitology. PaNic can be accessed at <http://purl.oclc.org/fishpest>.

The presence of a parasite on a host species is driven by evolution, biogeography, ecology, and chance (Esch et al. 1990, Combes 2001). Here, we present a web-tool for fish-parasite niche modeling (referred to as PaNic). PaNic is part of FishPEST (Fish Parasite Ecology Software Tools), which is a web project that integrates parasitological data with Fishbase (Froese and Pauly 2011). PaNic uses eco-biological parameters of the known hosts of a parasite species or genus as inputs to the Bioclim algorithm (Nix 1986), to compute niche boundaries of the parasite (Fig. 1). All candidate fish species are then evaluated for their compatibility with the parasite by comparing their ecological parameters with the computed niche boundaries. Note that, as used herein, the term 'niche' refers to variation in occurrence of the parasite in environmental space (Elith and Leathwick 2009). Although fairly related to realized niche, the present concept differs significantly from Hutchinson's definition (1957) because the variables used by PaNic represent only a subset of those possibly influencing parasite distribution on hosts. PaNic is implemented as a dynamic web system using the open source scripting language Python (van Rossum and de Boer 1991) and the Python-based web framework Django (Holovaty and Kaplan-Moss 2007).

Overview

A complete model can be built in less than a minute. Users 1) select the parasite species or genera they want to model; 2) choose weighted host ecological variables to be included in the computation of the parasite niche and; 3) narrow the list of candidate hosts (by fish family, locality or individual species). PaNic then searches online databases for known hosts of the parasite, determines the ecological characteristics that these hosts have in common, and then identifies other similar hosts that might support the parasite as well. Users with more detailed information can also develop custom models where they specify which hosts are known to be parasitized by a parasite (instead of relying on host lists available on the internet).

The results are organized in two parts. The first summarizes the main features of the model, comprising the selected parasite species, the list of known host species for the selected parasite, any known hosts that appear to be outliers of the set of known hosts, the overall phylogenetic similarity of the known hosts, eco-biological variables excluded from the model, and the lists of hypothesized 'likely' compatible, and 'less likely' compatible hosts. The second part contains, for each proposed host species, a graphical representation of the relative value of its ecological parameters with respect to the computed parasite niche (Fig. 2). A full summary of the model can be saved to a text file. Graphics are generated in pure HTML (i.e. they are not image files) and cannot be exported as single files. However, users interested in keeping the graphics can save the entire web page, or copy single graphics using a screenshot application. This simple option is convenient because hundreds of graphics can be stored in a relatively small HTML document.

Data

PaNic relies on an internal database including $>16\,000$ validated host/parasite records (for Acanthocephala, Cestoda,

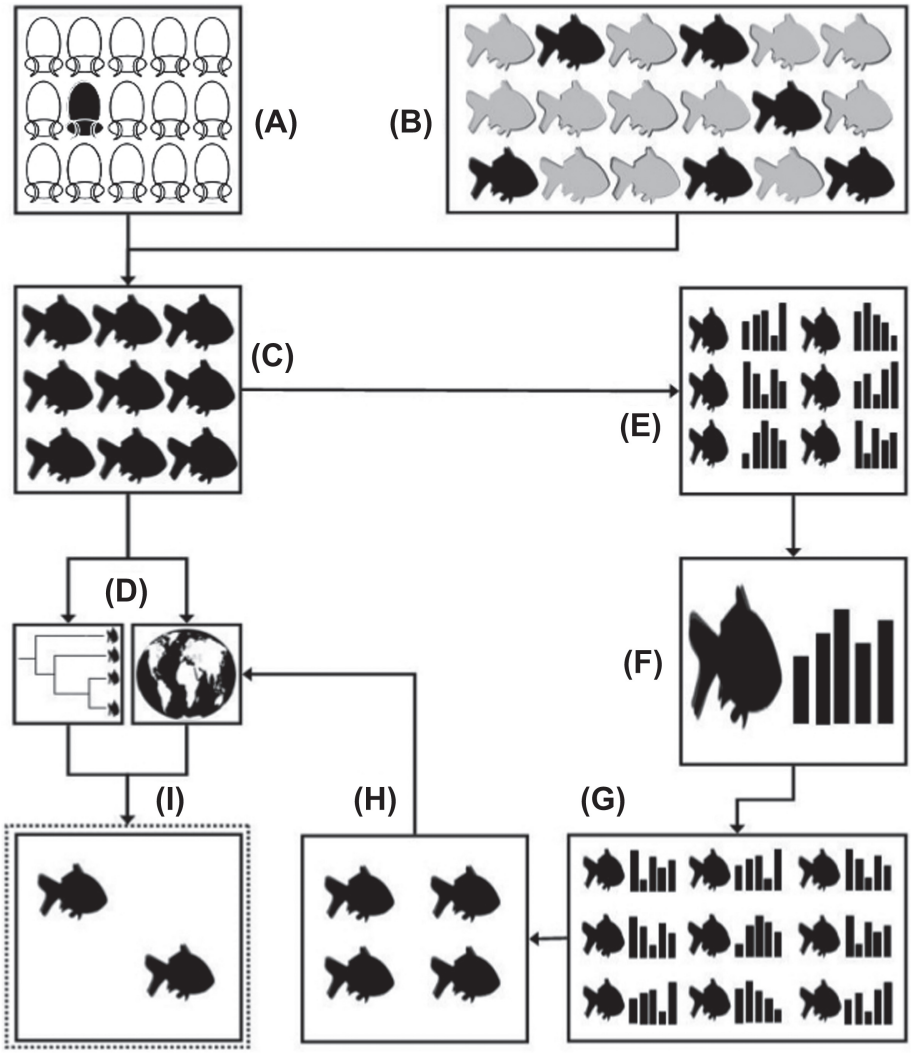

Figure 1. The steps PaNic takes to build a model. A known host set (C) is created by selecting a parasite species/genus (A), or by custom host selection (B); phylogenetic and biogeographical constraints of model hosts are set (D); eco-biological parameters of host set members (E) are used to create a niche model (F) for the selected parasite; the parasite niche model is then tested against a projection host set (G); members of the projection set with eco-biological parameters closest to niche model are individuated (H); after the application of a biogeographical and phylogenetic filter, the likely compatible hosts (I) are returned.

Monogenea, Nematoda, Trematoda) coming from scientific literature, internet databases, and museum collections (see PaNic online documentation for a detailed list of sources). Host names were validated according to Fishbase (Froese and Pauly 2011), while parasite names were validated according to the Catalogue of Life (Bisby et al. 2011) and the World Register of Marine Species (Appeltans et al. 2011). Only records (at species level) with valid scientific names or unambiguous synonyms for both host and parasite were retained. Invalid synonyms were replaced with accepted names only if unambiguous.

Although it would be good to include information about parasite life cycle stages, the data sources do not provide enough information to compile two separate host/parasite lists (adult vs larval stages). This is one of the reasons why PaNic contains a custom model option, where users interested in creating models for specific parasite life stages can integrate the information available from PaNic with other sources to compile a proper custom host list.

Fishbase provides Species Ecology Matrices that were used to compile an eco-biological matrix for > 27 400 fish species. Data in the Species Ecology Matrices were extracted using a script based on the Python HTML/XML parser Beautiful Soup (< www.crummy.com/software/BeautifulSoup/>). The Fishbase Species Ecology Matrices include > 15 ecological parameters, most of which are obtained from various combinations of the others (for specific details please refer to Fishbase documentation at < www.fishbase.org/manual/Key%20Facts.htm>). Of these, PaNic uses maximum length, growth rate (rate at which the asymptotic length is approached, termed K in Fishbase), life span, age at first maturity, and trophic level, which are independently measured and unbiased by autocorrelation. Fish size and trophic level are known to be important determinants of parasite communities (Sasal et al. 1999, Desdevises et al. 2002, Violante-González et al. 2010). The other three parameters were assumed relevant as they give complementary measures of individual metabolism and population resilience,

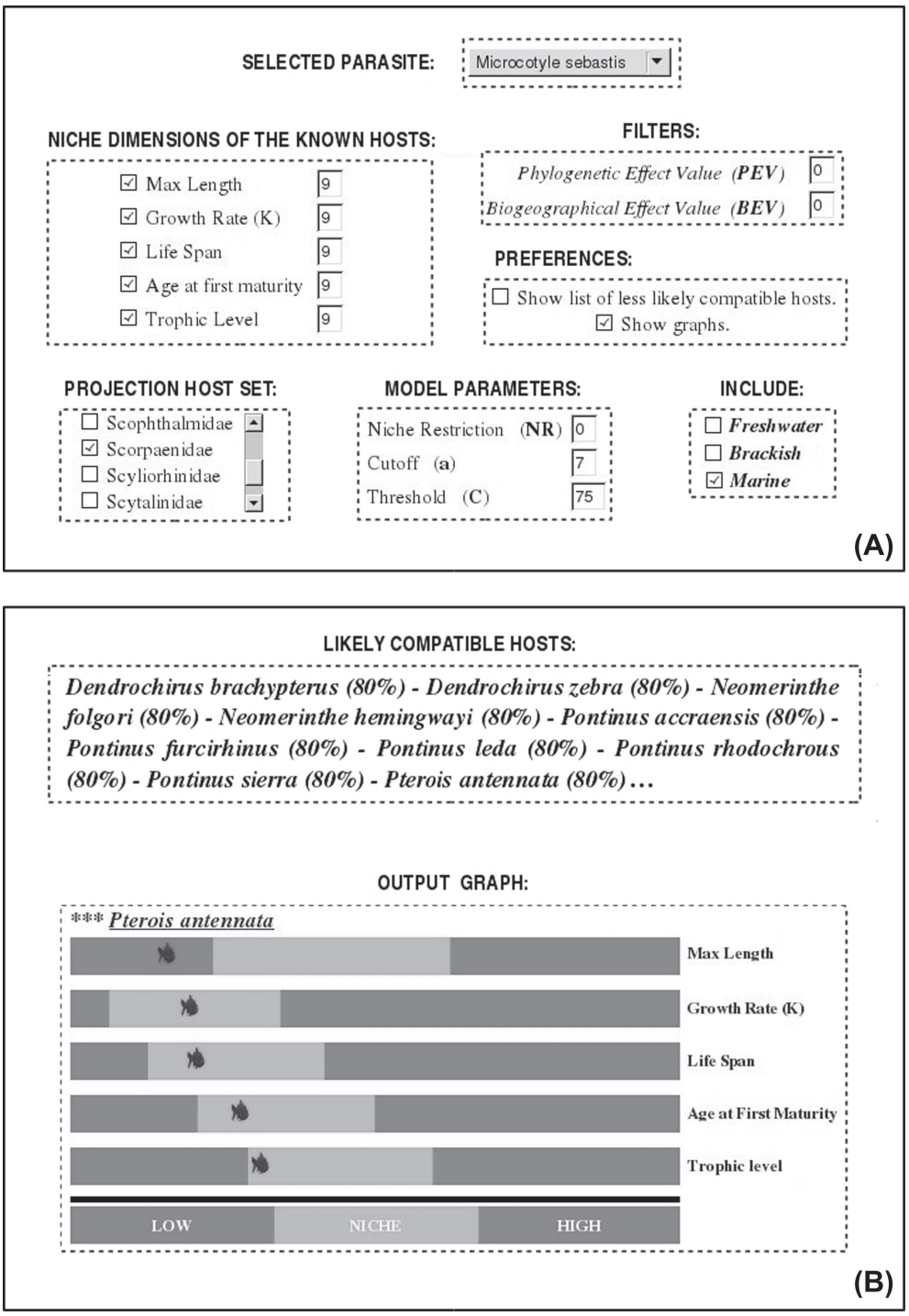

Figure 2. Example of PaNic input (A) and output (B). The model provides the first ten likely compatible hosts (among the Scorpaenidae) for the monogenean parasite *Microcotyle sebastis*. All parameters are set to default values. An example of graph for the likely compatible host *Pterois antennata* is presented; each horizontal bar represents the parasite compatibility interval for an ecological parameter; light grey portion of each bar represents the parasite niche interval for the parameter; fish symbols mark the relative position of each ecological parameter of *P. antennata* in respect to the computed niche for *M. sebastis*.

with potential influences on parasite infra- and component communities at ecological and evolutionary scales. For each species, habitat and geographical information available from Fishbase was included as well.

Choosing the right combination of variables may take trial and error. To increase the efficiency of PaNic we have chosen variables that are relatively independent. Despite the independent measure of the included variables, we cannot exclude eventual correlations among them (for example, between age at first maturity and life span, or between trophic ecology and max length). Still, potential correlations could vary substantially from one species to another, making it difficult to provide a general rule to determine the best performing combination of variables. Nevertheless, PaNic

provides a measure of relative variation of ecological variables (rVEV) that can be used as a guideline for to decide the parameters to include in model computation. A detailed explanation about how to use rVEV to improve model setup can be found on the online tutorial (that can be accessed from the Documentation page).

Known host set

If the user selects a parasite species or genus, PaNic populates a known host set, or KHS, (i.e. the set of known fish hosts for the selected parasite). Custom model section allows users to create models for parasite species not included in the list or for which the user has additional data on host use. To build a custom model, the user specifies a set of known host species for the parasite.

Outlier detection

After a KHS has been defined (both from the internal database, or through custom host selection), PaNic performs a procedure to detect outlier hosts that will be removed from the definition of the niche. Outlier hosts are not incorrect host/parasite records, so much as hosts in the KHS that differ significantly from the others in their eco-biological parameters. Differences are measured using Mahalanobis distance, which provide the standard test for outliers in multivariate data (Riani et al. 2009). Hosts are outliers if their average Mahalanobis distance from each other member of the KHS (computed on the basis of the considered eco-biological parameters) is larger than $(1 + NR \times 0.1)$ times the overall average Mahalanobis distance among each member of the set; *NR* (niche restriction value) is set to the maximum restrictive value of 0 by default, but users are allowed to select different values. The higher the value of *NR*, the fewer hosts will be included in the model.

Eco-biological parameters

Users can choose to include in the model any possible combination of 5 parameters (maximum length, K, life span, age at first maturity, and trophic level) and to attribute different weights (*W*) to each parameter. By default, all the parameters are included with an equal weight ($W = 9$). The inclusion of *W* in model computation is discussed in Parasite niche calculation section below.

Projection host set

The projection host set (PHS) is the set of potential hosts the user wants to test for compatibility with the selected parasite. By default, the PHS is empty. The PHS can be constrained by habitat (freshwater, brackish and marine), family, or locality. The projection host set can also be an explicit list (or single species) set by the user. For instance, if a user wants to know the likelihood that a particular parasite species or genus occurs in a particular species of fish, the user would select that fish species from the PHS check list.

Parasite niche calculation

Niche boundaries (*NB*) are calculated according to the Bioclim algorithm (Nix 1986). In most niche models, spatial correlations with environmental variables (e.g. temperature, rainfall) from known locations can be used to extrapolate to broader-scale distributions (Guisan and Thuiller 2005). PaNic modifies this technique to focus on host characteristics. Although there are alternative algorithms for ecological niche modeling (Elith et al. 2006), Bioclim offers the simplest way to combine up to 5 variables in PaNic. For each host eco-biological parameter (*E*), lower and upper *NB* are respectively calculated as:

$$mean(E) +/- (StDev(E) \times a \times 10^{-1}) \quad (1)$$

where *a* is a cutoff value that is set, by default, to 7 (according to Nix 1986). The higher the value of *a*, the wider the resulting niche intervals. For each parameter besides *NB*, compatibility boundaries (*CB*) are calculated as well. Lower and upper *CB* are calculated as the minimum and maximum of *E* values among the KHS.

For each member of the PHS, a high compatibility score (*hc*) is computed as:

$$\sum_{i,n}(Xi \times Ei \times Wi \times 100)/n \quad (2)$$

where *Ei* is the *i*-th *E*, *Wi* is the weight (*W*) assigned to *Ei*, and *n* is the number of *E*s included in the model. *Xi* is a binary operator indicating if *Ei* fits into *NB.* Similarly, a compatibility score (*c*) is computed with the same expression (3), with the only difference that *Xi* will indicate whether or not *Ei* fits into *CB.* A member of the PHS is considered likely compatible with the computed niche model if $hc > C$, where *C* is a threshold value which is by default set to 75; it is considered less likely compatible if $hc < C < c$.

Evolutionary and biogeographical effects

As already stated, evolution and biogeography may play a fundamental role in determining the distribution of parasite species among host species. PaNic determines which members of the PHS are most similar to those of the KHS for the selected parasite relative to selected eco-biological parameters. The results produced by PaNic should therefore be interpreted as a null model of the potential distribution of the parasite under the assumption that the selected eco-biological features are the major factors shaping parasite communities. In addition to providing potential host lists, users can test hypotheses about the relative importance of various phylogenetic and/or biogeographical effects on parasite assemblages.

Phylogenetic Effect Value (PEV)

The Phylogenetic Effect Value (PEV) weights the importance of evolutionary processes that might lead to host specificity. Parasites vary in host specificity and users can choose to consider the tendency for evolution when applying niche information. In PaNic, the potential effect of evolution can

be accounted for with a phylogenetic proximity index (*PPI*) that is calculated for each species of the PHS as:

$$(s + g + f)/3N \quad (3)$$

where *s*, *g* and *f* are the numbers of members of the KHS belonging, respectively, to the same species, genera and family, and *N* is the number of species in the KHS. The value of *s* is contained in *g* and *f*, and *g* is contained in *f* (but not vice-versa), providing a balance among different taxonomic levels, i.e. a known host species belonging to the same genera of the considered PHS member would give a contribution to PPI of (0 + 1 + 1)/3, while a known host species belonging to the same family but to a different genera would give a minor contribution of (0 + 0 + 1)/3. The greater the number of members of the KHS taxonomically close to the potential host species, the closer PPI is to 1. If there is no overlap even at the family level (i.e. there are no members of the KHS of the same family of the considered member of PHS), PPI = 0. The weight of a phylogenetic constraint can be included by increasing the value of *PEV* from 0 (default, no effect) to 9. A member of the PHS will be considered compatible to the selected parasite only when $PPI > PEV \times 0.01$.

According to Rhode's (1993) definition of host range, PaNic calculates a measure of phylogenetic proximity (*PP*) for the KHS, or:

$$(S + G + F)/3N \quad (4)$$

where *S*, *G* and *F* are, respectively, the number of members of the KHS whose species, genera and families are not unique to the KHS, and *N* is the size of the KHS. *PP* is maximum (= 1) when all members of the KHS are subspecies of the same species. It will be 0 when each member belongs to a different family. High values of *PP* (> 0.5) should encourage users to account for evolutionary effects in their model, by increasing *PEV*.

Biogeographical Effect Value (BEV)

The Biogeographical Effect Value (*BEV*) weights the importance of geography on parasite distributions. For instance, some parasites are cosmopolitan, while others are restricted to oceans or latitudinal ranges, islands, or continental regions. Locality records for the fish of the KHS are compared to those of each member of the PHS. For each member of the PHS, a Biogeographical Compatibility Index (*BCI*) is calculated as:

$$sl/L \quad (5)$$

where *sl* is the number of localities where the fish species occurs together with some other member of the KHS, and *L* is the total number of localities for KHS. The weight of a biogeographic constraint can be included by increasing the value of *PEV* from 0 (default, no effect) to 9. A member of the PHS will be included in list of compatible hosts only if $BCI > BEV \times 0.1$, independently from its ecological features.

An approximate indication of biogeographical range (*BR*) within the KHS, is calculated as:

$$L/Lt \quad (6)$$

where *Lt* is the total number of locality records for the KHS (including repetitions), and *L* is the number of localities where at least one member of the KHS occurs (which is equal to *Lt* with no repetitions). *BR* is maximum (1) if the geographical distributions of the KHS are non-overlapping. Low values of *BR*, suggest the parasite has a limited biogeographical range, arguing for users to increase *BEV*.

Potential applications

PaNic is a useful application for summarizing existing information. It can easily create host lists for parasite species or genera. PaNic can provide standardized measures of the extent of host range or geographic distribution of fish parasites, allowing a range of hypotheses to be tested. It can indicate which known hosts are outliers in terms of maximum length, growth rate (K), life span, age at first maturity or trophic level, providing insight into host-parasite evolution and suggesting potential cryptic parasite species. Perhaps the most creative feature of PaNic is that it provides potentially suitable hosts for a parasite based on ecological, geographical, and phylogenetic information. This could be advantageous when trying to determine, for instance, the potential host range of an introduced parasite. PaNic could also be useful when initiating parasitological studies on a previously unexamined or underexamined fish species. Although PaNic does not provide a confidence interval per se, users can choose a tradeoff between type-I and type-II error in the estimate by broadening or narrowing criteria for including potential hosts.

To cite PaNic or acknowledge its use, cite this software note as follows, substituting the version of the application that you used for 'version 6':

Strona, G. and Lafferty, K. D. 2012. How to catch a parasite: Parasite Niche Modeler (PaNic) meets Fishbase. – Ecography 35: 481–486 (ver. 6).

Acknowledgements – Any use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the US government. The authors would like to thank Rainer Froese and Nicolas Bailly for their suggestions and support with Fishbase data.

References

Appeltans, W. et al. 2011. World Register of Marine Species. – <www.marinespecies.org>, accessed 23 May 2011.

Bisby, F. A. et al. 2011. Species 2000 & ITIS Catalogue of Life: 2011 annual checklist. – <www.catalogueoflife.org/annual-checklist/2011/>, accessed 23 May 2011.

Combes, C. 2001. Parasitism: the ecology and evolution of intimate interactions. – Univ. of Chicago Press.

Desdevises, Y. et al. 2002. Evolution and determinants of host specificity in the genus *Lamellodiscus* (Monogenea). – Biol. J. Linn. Soc. 77: 431–443.

Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Esch, G. W. et al. 1990. Parasite communities: patterns and processes. – Chapman and Hall.
Froese, R. and Pauly, D. 2011. FishBase. – <www.fishbase.org>, accessed 23 May 2011.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Holovaty, A. and Kaplan-Moss, J. 2007. The definitive guide to Django: web development done right. – Apress.
Hutchinson, G. E. 1957. Concluding remarks. – Cold Spring Harbor Symp. Quant. Biol. 22: 415–427.
Nix, H. A. 1986. A biogeographic analysis of Australian elapid snakes. – In: Longmore, R. (ed.), Atlas of Elapid snakes of Australia. Australian Flora and Fauna Series no. 7, pp. 4–15.
Riani, M. et al. 2009. Finding an unknown number of multivariate outliers. – J. R. Stat. Soc. B 71: 1–20.
Rohde, K. 1993. Ecology of marine parasites, 2nd ed. – CAB International.
Sasal, P. et al. 1999. Specificity and host predictability: a comparative analysis among monogenean parasites of fish. – J. Anim. Ecol. 68: 437–444.
van Rossum, G. and de Boer, J. 1991. Interactively testing remote servers using the Python programming language. – CWI Q. 4: 283–303.
Violante-González, J. et al. 2010. Factors determining parasite community richness and species composition in black snook *Centropomus nigrescens* (Centropomidae) from coastal lagoons in Guerrero, Mexico. – Parasitol. Res. 107: 59–66.