ISSN 0013-8738, Entomological Review, 2015, Vol. 95, No. 2, pp. 285–288. © Pleiades Publishing, Inc., 2015.
Original Russian Text © A.L. Lobanov, 2015, published in Entomologicheskoe Obozrenie, 2015, Vol. 94, No. 1, pp. 197–202.

The Diagnostic Value of Qualitative and Quantitative Characters in Computer Identification Keys

A. L. Lobanov

Zoological Institute, Russian Academy of Sciences, St. Petersburg, 199034 Russia
e-mail: all@zin.ru

Received May 12, 2014

Abstract—A new method for estimating the diagnostic value of characters for computer keys to biological objects is proposed. The algorithm described can be used for correct comparison of qualitative and quantitative characters. A brief review of the history of biological diagnostics is given.

DOI: 10.1134/S0013873815020128

Biological diagnostics, i.e., the branch of taxonomy studying the theory and practice of creating keys to biological objects, has been in existence for half a century. The initial period of its development, which started with the works of Russian researchers (Balkovsky, 1962; Kiskin et al., 1965), coincided with the time when computers became available to biologists. Optimization of keys with the help of computers facilitated rapid progress of diagnostics and resulted in a number of working implementations (Goodall, 1968; Morse, 1968; Lobanov, 1974; Dallwitz, 1974; Pankhurst, 1975, 1978). Since the end of the XX century, all the noteworthy advances in biological diagnostics have involved the use of computers (Lobanov and Dianov, 1994; Smirnov et al., 1996; Dianov and Lobanov, 1997, 1999; Lobanov, 1997, 2007; Ryss and Lobanov, 1999; Lopatin and Dovgailo, 2002; Dmitriev, 2006; Vakhitov et al., 2007; Kirejtshuk et al., 2011).

The term "computer-aided biological identification" presently refers to a fairly broad concept including essentially different ways of statement and solution of diagnostic problems (Lobanov and Ryss, 1999). Below we will consider only the traditional computer-aided systems of taxonomic diagnostics, in which the taxa and their characters are defined by the author; the user evaluates the characters of the given plant or animal specimen and enters them into the computer program, which then attempts to assign this specimen to one of the preset taxa. Most programs developed by biologists belong to this type. The computer diagnostic systems implementing automatic image processing and those based on the image discrimination theory are beyond the scope of this paper.

Our classification of the types of biological keys was developed long ago (Lobanov, 1972) and remained almost unmodified since that time. Only the meaning of the term "character" changed. Earlier, a distinct element of description of a biological object was referred to as a "class of characters," and its particular states, as "characters"; now, the researchers consider a "character" which can have different "character states." In addition, the modern computer keys can be classified into single-user programs running on a local computer and multiple-user programs running on a server and accessible via the Internet. Most computer keys are multi-entry and polytomous. One more important characteristic of computer keys is the number of possible states of a particular character in one taxon, according to which the key could be unimodal or polymodal.

Despite the external differences in the interface, graphics, and details of the user dialog organization, the numerous diagnostic programs written by different authors implement virtually the same principal algorithm which has been fully perfected in terms of logic and utility. At the beginning of the session, the user selects the initial set of the taxa to one of which the object in question will be eventually assigned. Advanced diagnostic systems make use of the known hierarchy of taxa in this set and may give the user an option of choosing the required level of identification (for example, the key may contain species from many genera but the user may only need to identify a given object to genus). Then, the system usually allows the user to select the preferable subset of characters and the order of their presentation. The rest of the work consists of several identical steps of diagnosis. At the

beginning of each step, the user is given a set of characters suitable for diagnosis, from which one or several characters can be selected; then the user considers the states distinguished for each character, compares these states with the properties of the object in question, and enters the information on the matching states into the program. The system then selects the taxa possessing the matching character states for the next step, and re-evaluates the remaining characters. Such steps are repeated until only one taxon is left, which means that identification is completed; the available information on this taxon is then presented to the user.

An important characteristic of a computer-aided diagnostic system is the way of presenting characters at any step of diagnosis. Characters may be presented in a fixed order, for example, alphabetically or according to some rules defined by the author. However, in advanced systems, the characters are reordered at each step according to their importance for differentiation among the remaining taxa. This estimate of the usefulness of a particular character, changing from one step of diagnosis to the next, is called the diagnostic value. Several formulas have been proposed for calculating the diagnostic value (Pankhurst, 1970; Lobanov, 1974; Gambaryan, 1975). In the author's dissertation (Lobanov, 1983), six different formulas were compared using the data of 16 actual multi-entry keys to plants and animals with the number of taxa varying from 8 to 73. Although different at first sight, all these formulas were based on determining the number of states of each character in the taxon-character matrix (either the initial complete matrix or the reduced matrix for the taxa remaining at the second and subsequent steps). The formulas were evaluated by a very simple algorithm. The taxon-character matrix of a given key was entered into the computer, and a special diagnostic program imitated identification of each taxon included in that key. At each step, the characters were ranked by the formula being tested, and the character with the highest diagnostic value was used in diagnosis. The total number of steps required for complete identification of all the taxa was divided by the number of taxa in the key to calculate the mean length of the identification path for each formula. Among others, our original formula was tested:

$$d = \frac{N^2}{\sum_{i=1}^{s} n_i^2},$$

where d is the diagnostic value, N is the total number of taxa at a given step, s is the number of distinct states of the character in question, and n is the number of taxa possessing the character state i. This formula proved to be either as good as, or better than the others. It was found to be slightly inferior in only one case (a key with 19 taxa; the mean path length 2.79 against 2.74 by a different formula); therefore, in our subsequent work we used only this formula.

The comparison of formulas carried out nearly 40 years ago was mostly based on the material of unimodal keys (i.e., keys in which only one state of a character was specified for each taxon). However, in case of polymodal keys, in which the number of character states in a taxon was not limited and may be as great as s, the use of this formula could lead to error when only two taxa were left at the last step, so that the programs had to be corrected. Besides this drawback, the critical weakness of this formula lies in the fact that it can be applied only to qualitative characters with a limited number of distinguishable states. According to the recommendations of psychologists (Miller, 1956), no more than 7–9 distinct states of one character are usually considered, since the short-term memory of man has been shown to be limited by 7 ± 2 comparable information blocks. Until recently, we had to transform metric and meristic characters (body length, length-to-width ratio of some body part, number of segments of some organ, etc.) into qualitative ones by splitting the possible range of their values into fixed intervals which were then interpreted as "character states." This approach was inconvenient for both the author and the user of the key, who had to select two intervals if the actual value fell on the boundary between them.

There is an obvious way to avoid this difficulty: the key should include quantitative characters represented in the taxon-character matrix by two values (the minimum and the maximum) for each taxon. If the characters are presented to the user in a fixed order, this approach poses no problem. However, if the characters are to be ranked by their diagnostic value, one has to obtain comparable estimates of the values of qualitative and quantitative characters. Below, we propose a new method of calculating the diagnostic value of characters, which allows qualitative and quantitative characters to be used simultaneously and on equal terms.

The algorithm of character evaluation for the current set of taxa (at the current step of diagnosis) can be easily described. All the characters available at the

current step are enumerated (if the step is not the first one, the characters already used are discarded). The taxon-character matrix is analyzed for each character. The remaining taxa are compared pairwise in all the possible combinations. If the set of states of a qualitative character in one taxon shares at least one state with the corresponding set in the other taxon (or if the ranges of a quantitative character adjoin or overlap in the two taxa), this situation is considered as a match; otherwise, a mismatch is recorded. The sum of mismatches for the given character is divided by the sum of matches plus one (to avoid division by zero if no matches are present). This quotient is taken as an estimate of the diagnostic value of the character.

An obvious advantage of this method is complete comparability of the estimates for qualitative and quantitative characters. This simplifies the diagnostic programs and allows the characters to be grouped according to biologically significant aspects. To demonstrate the practical advantage of the new method, we compared it with the old formula by the algorithm described above, using the dataset of the actual key to beetles (Lobanov and Dianov, 1996) with 130 taxa and 24 characters. The mean length of the identification path was 5.36 for the old formula and 5.15 for the new method, which thus yielded a slight but tangible gain.

One of the drawbacks of the new method is a greater amount of computations. The old formula uses N retrievals from the taxon-character matrix whereas the new method requires $N \times (N–1)/2$ retrievals, i.e., approximately $N/2$ times as many. However, the constantly increasing computational power of modern processors makes this limitation negligible.

The new algorithm is difficult to express by a single formula. We have attempted to describe it using the set theory notation:

$$d = \frac{\sum\limits_{i=1,\, j=i+1}^{N-1,\, N} (T_i \cap T_j = \varnothing)}{1 + \sum\limits_{i=1,\, j=i+1}^{N-1,\, N} (T_i \cap T_j \neq \varnothing)}$$

Besides the designations explained above, this formula includes T, which is the set of states of a given character (or the set of possible numerical values of a given quantitative character) present in a given taxon. The numerator contains the sum of cases when the intersection of the taxon-character matrix points for the two taxa is empty; the denominator, the sum of cases when this intersection is not empty. The character with a greater sum of empty intersections will obviously have a greater value for diagnostic purposes.

REFERENCES

1. Balkovsky, B.E., "On the Ways of Increasing the Diagnostic Value of Characters Used for Identification of Plants," Botan. Zh. **47** (9), 1309–1314 (1962).
2. Dallwitz, M.J., "A Flexible Program for Generating Identification Keys," Syst. Zool. **23** (1), 50–57 (1974).
3. Dianov, M.B. and Lobanov, A.L., "PICKEY, a Program for Organism Identification with Interactive Use of Images," Trudy Zool. Inst. Ross. Akad. Nauk **269**, 35–39 (1997).
4. Dianov, M.B. and Lobanov, A.L., "BIKEY8—the Biological Diagnostic Software for Windows," Trudy Zool. Inst. Ross. Akad. Nauk **278**, 74 (1999).
5. Dmitriev, D.A., "3I, a New Program for Creating Internet-Accessible Interactive Keys and Taxonomic Databases and Its Application for Taxonomy of Cicadina (Homoptera)," Rus. Entomol. J. **15** (3), 263–268 (2006).
6. Gambaryan, P.P., "A Numerical Key to the Aquatic Flowering Plants of Armenia," Biol. Zh. Armen. **28** (9), 108–111 (1975).
7. Goodall, D.W., "Identification by Computer," BioScience **18** (6), 485–488 (1968).
8. Kirejtshuk, A.G., Lobanov, A.L., Smirnov, I.S., et al., "Internet-Based Keys to Biological Objects: Five Years Later," in *Internet Scientific Services: the Exaflop Future: Proc. of Int. Conf. on Supercomputers (September 19–24, 2011, Novorossiysk)* (Moscow State Univ., Moscow, 2011), pp. 449–453.
9. Kiskin, P.Kh., Pecherskaya, I.N., and Pechersky, Yu.N., "Automated Diagnostic Search for Grape Varieties Using a Minsk-1 Computer," Vinodel. Vinograd. SSSR, No. 1, 21–22 (1965).
10. Lobanov, A.L., "Logical Analysis and Classification of the Existing Types of Diagnostic Keys," Entomol. Obozr. **51** (3), 668–681 (1972).
11. Lobanov, A.L., "Estimation of the Diagnostic Value of Character Series in Computer-Oriented Multi-Entry Keys," in *Abstracts of Papers, VI Conf. of Young Sciences of Komi Republic* (Syktyvkar, 1974), pp. 125–126.
12. Lobanov, A.L., *Candidate's Dissertation in Biology* (Leningrad, 1983).
13. Lobanov, A.L., "Computer Keys in Biology: the Results of 30 Years of Evolution," in *Computer Databases in Botanical Research: Collected Papers* (St. Petersburg, 1997), pp. 51–55 [in Russian].
14. Lobanov, A.L. and Dianov, M.B., "The Interactive Computer Diagnostic System BIKEY and the Possibili-

ties of Its Use in Entomology," Entomol. Obozr. **73** (2), 465–478 (1994).
15. Lobanov, A.L. and Dianov, M.B., *The World of Beetles [Wir Bestimmen Käfer]: a CD-ROM and a Brief Manual* (Dialobis Edition, Berlin, 1996).
16. Lobanov, A.L. and Ryss, A.Yu., "Computer Identification Systems in Zoology and Botany: the Present State and Prospects," Trudy Zool. Inst. Ross. Akad. Nauk **278**, 17–19 (1999).
17. Lopatin, I.K. and Dovgailo, K.E., *Leaf Beetles of the Genus Cryptocephalus (Chrysomelidae) of the Palaearctic: a Key to 398 Species and a Lysandra Database on a CD-ROM* (Minsk, 2002).
18. Miller, G.A., "The Magical Number Seven, Plus or Minus Two," Psychol. Rev. **63**, 81–92 (1956).
19. Morse, L.E., "Construction of Identification Keys by Computer," Amer. J. Botan. **55** (6), 737 (1968).
20. Pankhurst, R.J., "A Computer Program for Generating Diagnostic Keys," Computer J. **13** (2), 145–151 (1970).
21. Pankhurst, R.J., *Biological Identification with Computers* (Academic Press, London, 1975).
22. Pankhurst, R.J., *Biological Identification. The Principles and Practice of Identification Methods in Biology* (Edward Arnold, London, 1978).
23. Ryss, A. and Lobanov, A., "Principles of Taxonomic Identification Illustrated on Nematode Computer Key," Proc. Zool. Inst. Russ. Acad. Sci. **280**, 22–23 (1999).
24. Smirnov, I.S. and Lobanov, A.L., "The Computer Key to Brittle Stars as a Taxonomic Database," Byul. Mosk. O-va Ispyt. Prir. Otdel. Geol. **72** (1), 87–88 (1999).
25. Smirnov, I., Lobanov, A.L., and Dianov, M.B., "Creation of Computer Picture Identification Key for the Arctic Ophiuroids," in *Program and Abstracts of the 9th Int. Echinoderm Conf., August 5 to 9, 1996, San Francisco, California* (San Francisco, 1996), p. 132.
26. Vakhitov, A.T., Granichin, O.N., Kirejtshuk, A.G., and Lobanov, A.L., "Minimum-Length Questionary Algorithms for an Internet-Based Biological Key and Their Implementation," in *Internet Scientific Services: Proc. of All-Russian Conf. Dedicated to the 15th Anniversary of the Russian Foundation for Basic Research (September 24–29, 2007, Novorossiysk)* (Moscow State Univ., Moscow, 2007), pp. 293–295.