Biodiversity Informatics, 9, 2014, pp. 1–12

CHARACTER SELECTION DURING INTERACTIVE TAXONOMIC IDENTIFICATION: "BEST CHARACTERS"

*NADIA TALENT[1], RICHARD B. DICKINSON[1], TIMOTHY A. DICKINSON[1,2]

[1]*Department of Natural History, Royal Ontario Museum, 100 Queen's Park, Toronto, M5S 2C6, Canada*

[2]*Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, M5S 3B2, Canada*

Corresponding author: Nadia Talent, nadia.talent@utoronto.ca

Abstract— Software interfaces for interactive multiple-entry taxonomic identification (polyclaves) sometimes provide a "best character" or "separation" coefficient, to guide the user to choose a character that could most effectively reduce the number of identification steps required. The coefficient could be particularly helpful when difficult or expensive tasks are needed for forensic identification, and in very large databases, uses that appear likely to increase in importance. Several current systems also provide tools to develop taxonomies or single-entry identification keys, with a variety of coefficients that are appropriate to that purpose. For the identification task, however, information theory neatly applies, and provides the most appropriate coefficient. To our knowledge, Delta-Intkey is the only currently available system that uses a coefficient related to information theory, and it is currently being reimplemented, which may allow for improvement. We describe two improvements to the algorithm used by Delta-Intkey. The first improves transparency as the number of remaining taxa decreases, by normalizing the range of the coefficient to [0,1]. The second concerns numeric ranges, which require consistent treatment of sub-intervals and their end-points. A stand-alone Bestchar program for categorical data is provided, in the Python, R, and Java languages. The source code is freely available and dedicated to the Public Domain.

Key words— Separation coefficient, polyclave, multi-access key, entropy, Delta-Intkey, information theory

In biodiversity informatics, one type of automated tool for taxon identification is the polyclave, also called a multiple-entry key or multi-access key, an information retrieval system (Duke 1969) that sequentially accepts specifications of character-states observed, in any order (Morse 1975; Duncan and Meacham 1986; Pankhurst 1991), and "eliminate[s] taxa which disagree with the specimen to be identified" (Pankhurst and Aitchison 1975). Some confusion surrounds these and related terms (Gower 1975; Hagedorn et al. 2010). A different type of system that is frequently said to "identify" (Pankhurst 1975; MacLeod 2007) establishes groupings of known and/or unknown individuals by using similarity or dissimilarity measures; thus, it establishes the taxa that form the basis of the dataset used by a polyclave. This is "clustering", or in the terminology of Sneath and Sokal (1973, p. 3), "classification", rather than "identification". Sneath and Sokal, whose terminology we follow, define identification (p.3) as "the allocation or assignment of additional unidentified objects to the correct class once a classification has been established." Software tools for classification are sometimes packaged together with tools for building polyclaves (e.g., ETI Bioinformatics undated).

The long-standing tradition for printed keys, since de Lamarck's development of these tools (de Lamarck 1778), is to define a key as "an artificial analytical device or arrangement whereby a choice is provided between two contradictory propositions" (Voss 1952), thus requiring binary characters. The character set and character states are usually carefully chosen and limited in number to save space. Considerable effort is required to choose the characters and build such keys (Hagedorn et al. 2010). The characters are often chosen for ease of use, or else to form "diagnostic descriptions" ("irredundant character sets") that concisely differentiate every taxon (reviewed by

Payne and Preece 1980). Usually, the characters chosen do not vary within a taxon. It is common, but unnecessary, to be consistent with a natural taxonomy, and it may be undesirable to limit the character selection in that way (Voss 1952). A polyclave, however, may hold quite different data, possibly with multi-state characters, using characters that vary within a taxon, and including states coded as "missing data" for all taxa.

Practice has shown that polytomies in printed identification keys can lead to considerable confusion (Voss 1952), and they are almost universally avoided in that context (Hagedorn et al. 2010). Many of the identification keys that are being made available online also either use binary characters only (Toda et al. 2004; Guala 2004–12; Christensen 1999; Rosatti undated) and/or are automated versions of traditional keys (Martellos 2010) called "pathway keys" by Walter and Winterton (2006). Software is available, however, for implementing general polyclaves that allow multi-state characters and an unconstrained sequence of character selection (Dallwitz 1993; Brach and Song 2005; Alexander 2006; Lucidecentral.org 2010; Ung et al. 2010).

A polyclave can be difficult to use correctly, such as by novices who make frequent errors, or for polythetic taxa which are distinguished by possessing a majority of a set of character states, rather than by required states (Morse 1971, 1975). These problems can be treated with polythetic polyclaves that provide an "error-tolerance" or "mismatch threshold" setting, and by weighting character states and taxa (specifying which character states or taxa are rare). Here we consider those features to be secondary additions to a simple interface that allows character and character-state selection in any order. We concentrate on an efficient error-free identification process in a monothetic key.

In both printed keys (Walters 1975; Lobanov et al. 1981) and polyclaves, permitting multi-way choices (polytomies, multi-state characters) can gain considerable efficiency in the sense that the expected number of choices required to arrive at an identification is greatly reduced (Cover and Thomas 2006). Ideal characters for speeding up the identification are those that separate the taxa evenly into small groups (Osborne 1963; Lobanov et al. 1981; Cover and Thomas 2006). Quite complex characters with overlapping states (including numeric ranges; Dallwitz et al. 2002) can have good separating power.

Some of the software that provides a polyclave interface calculates a coefficient to rank the characters by their separating power, to guide the user to an efficient identification of an unknown specimen. The coefficients have names such as "Best character" (Dallwitz et al. 1998/2012), "Separation coefficient" (ETI Bioinformatics undated), "dichotomizing value" (Morse, 1971), or "Best" (Brach and Song 2005; Pankhurst, 1978; Lucidecentral.org 2010). Other software provides a choice of coefficients (Ung et al. 2010), such as variants of a Jaccard coefficient, or Simple Matching coefficient (also called "Sokal & Michener coefficient", Sneath and Sokal 1973). Those ranking coefficients are different from, and complementary to, a "differentiating characters" list (e.g., as provided in MEKA; Duncan and Meacham 1986; Duncan and Meacham 1987; Christensen 1999; Rosatti undated), which lists those (binary) character states that uniquely identify a single taxon. Some systems use the same coefficients both for assessing a proposed taxonomy and for designing identification keys (ETI Bioinformatics undated). It should be emphasized that the coefficients used in polyclaves and their terminology are related to, but are used in a very much simpler setting than, various techniques in multivariate data exploration. The many techniques grouped under headings such as "character ranking" (Podani 2000) or "feature selection" or "weighting" (Dale et al., 1986), evaluate the contribution of characters to the most significant patterns present, as a preparation for reducing the dimensionality of the data. With those methods, characters are considered all together or in pairs, whereas for a polyclave the characters are assessed individually, producing a single numeric value for each.

Pure polyclave software aims only to provide a hint for the user, a set of numeric values that indicate how well each character differentiates the remaining taxa. Although different types of identification software and embellishments to polyclave software have been constructed that take taxon probabilities and character-state probabilities into account, expressed as weighting schemes and conditional characters (called "controlling" and "dependant" characters in DELTA; Dallwitz et al. 2010), those probabilities often cannot be known (reviewed by Pankhurst, 1978). The user working with a particular specimen is often the only meaningful source of an assessment of a character's usefulness, because it depends on whether it is feasible to assess the character state at all. The usefulness of a character ranking probably amounts only to indicating which characters are very good and which very poor at differentiating; nonetheless, correct calculation is needed to align the characters appropriately.

Here we review the calculation of the character-ranking coefficient derived from information theory and recommend its use for polyclaves. It is similar to but not the same as Delta-Intkey's "Best character" function (Spooner and Chapman 2007; Dallwitz 2010). Statements made here about the Best Character implementation in Delta-Intkey are based on our reverse engineering. Dallwitz (1974) gives a formula that is similar to Delta-Intkey's behaviour if one interprets "*n* sub *j*, the number of taxa in the *j*th subgroup" as a proportion.) Information-theoretic character ranking can be counter-intuitive for certain types of characters, and has never, to our knowledge, been fully implemented for taxonomic applications.

CHOICE OF COEFFICIENT

Formulae for character-ranking coefficients fall into two types: (i) a separation number or separation coefficient reflects the number of groups that can be distinguished by using the character, whereas (ii) the Information statistic or entropy *H* reflects both the number of groups and the number of taxa in each, that is, the evenness of the division (Cover and Thomas 2006).

The separation number is a count of the number of pairs of taxa that can be distinguished using the character (Table 1). The separation coefficient is the separation number divided by the number of possible pairs, so it is normalized to the

Table 1: Various separation measures for characters in a simple constructed key. '/' in a list of character states indicates 'or'; *n* is the total number of states; bold-faced values are less informative than the corresponding values of some other coefficients.

	Anther color	**Petal color**	**Stamen #**	**Style #**
Taxon 1	red	red	10	2
Taxon 2	pink	pink/white	10	5
Taxon 3	white	white	20	5
Taxon 4	yellow	yellow	20	5

					Maximum value
Separation number	6	5	4	3	$n(n-1)/2$
Separation coefficient	1	0.83	0.67	0.50	1
Simple matching (Xper2, observed)	1	**1**	0.67	0.50	N/A
Jaccard (Xper2, observed)	1	**1**	0.67	0.50	N/A
Pairwise average Jaccard distance	1	0.92	0.67	0.50	1
Information (in bits)	2	1.63	1	0.81	$\log_2 n$
Normalized to the range [0,1] for 4 taxa	1	0.81	0.50	0.41	1

Table 2: Taxa and character states can be specified as a joint probability distribution, by assuming that taxa are equiprobable, and that alternative states are equiprobable for a taxon. '/' in a list of character states indicates 'or'.

	Flower color
Taxon 1	white
Taxon 2	orange/pink
Taxon 3	pink/red

	white	**orange**	**pink**	**red**	
Taxon 1	1/3	0	0	0	1/3
Taxon 2	0	1/6	1/6	0	1/3
Taxon 3	0	0	1/6	1/6	1/3
	1/3	1/6	1/3	1/6	

range [0,1]. If the separation coefficient takes the value 1, this signals that the answer to a single question about the state of that character would be enough to distinguish every taxon, a feature that we believe could be helpful to the polyclave user, particularly when a large number of possible identities remain.

Most authors who have used this approach have excluded any taxa that overlap, i.e., that share some but not all character states (e.g., Table 1, petal color = white exhibits overlap). Morse (1971) extended the separation number to count the extent of overlap, but only for binary characters. The Simple Matching ("Sokal & Michener") coefficient, the ratio of character-state matches to character-state pairs, and the Average Pairwise Jaccard Distance (which excludes "negative" , also called "absence" matches, i.e., shared not-applicable states) are similar, relatively simple, calculations. Variations on these coefficients have also been advocated (e.g., the Jaccard coefficient of the Xper2 system of Ung et al. 2010). If overlapping states are fully incorporated, then the coefficient becomes the same as the Information content. When designing a pathway key, state-overlap is best avoided if other characters are available that cleanly distinguish the taxa, and a separation coefficient shows those characters to advantage while the information content does not.

THE INFORMATION COEFFICIENT

Shannon's information theory (Shannon 1948) has long been used for problems similar to the taxon-identification problem (e.g., Shwayder 1971, 1974), and is often cited as an appropriate foundation for ranking the characters in a polyclave (Pankhurst 1991). Although it has been stated that "the coefficient is not defined for continuously-varying numerical data" (Lance and Williams 1966), this is not relevant for polyclaves because the identification database will treat numeric ranges in most respects like other discrete categories (see below).

If an event e occurs with probability $p(e)$, and we are told that e has occurred, then we have received:

$$I(e) = \log\frac{1}{p(e)} = -\log p(e) \qquad \textbf{(equation 1)}$$

units of information (Abramson 1963). For example, if one of two equally likely events is specified, then one *bit* of information is obtained. The information content is a lower bound on the number of yes/no questions that will lead to each of the possible identifications (Cover and Thomas 2006 chapter 5). For a single question, equivalent to each individual character in a polyclave, the mutual information $I(X;Y)$ is "the reduction in uncertainty of X due to the knowledge of Y", where X is the taxon identity and Y is the taxonomic character:

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad \textbf{(equation 2)}$$

where $x \in X$ and $y \in Y$ are the possible values of random variables X and Y

Entropy (H) "is the minimal descriptive complexity of a random variable", and "mutual information is the relative entropy between the joint distribution and the product distribution" (Cover and Thomas 2006). For the example in table 2 see figure 1.

Figure 1: Calculation of the Information content for the example in Table 2

$$I(X;Y) = \frac{1}{3}\log\left(\frac{\frac{1}{3}}{\frac{1}{3}\times\frac{1}{3}}\right) + \frac{1}{6}\log\left(\frac{\frac{1}{6}}{\frac{1}{3}\times\frac{1}{6}}\right) + \frac{1}{6}\log\left(\frac{\frac{1}{6}}{\frac{1}{3}\times\frac{1}{3}}\right) + \frac{1}{6}\log\left(\frac{\frac{1}{6}}{\frac{1}{3}\times\frac{1}{3}}\right) + \frac{1}{6}\log\left(\frac{\frac{1}{6}}{\frac{1}{3}\times\frac{1}{6}}\right)$$

A succinct interpretation of how the formulae apply to a polyclave is given by Pankhurst (1991). If $p1$, $p2$, $p3$, … , pm are the proportions of states 1, 2, 3, … , m for a group of taxa, then "the 'information' that we get from seeing state 1 on a specimen is the effect this fact has on our opinion of what taxon we think we have. If $p1$ was 1 (i.e. all taxa show character state 1 only), then on seeing state 1 we would gain no information at all."

NORMALIZING THE COEFFICIENT

In equations 1 and 2, the choice of the base for the logarithm is arbitrary, and the units of information are called *bits* if logarithms to base 2 are used, *Hartleys* (Abramson 1963) with base 10, and *nats* (Cover and Thomas 2006) (or *nits* (MacDonald 1952)) if natural logarithms are used. For the example in table 2:

$$I(X;Y) = \tfrac{2}{3}\log(3) + \tfrac{1}{3}\log\left(\tfrac{3}{2}\right) \approx 1.251 \text{ bits, } 0.38 \text{ nats}$$

The entropy H of a character takes a maximum value when the probabilities are uniformly distributed (Cover and Thomas 2006, Theorem 2.6.4), and that maximum is log n where n denotes "the number of elements in the range". This result has been stated as "maximal H… depends only on the number of states" (Abbott et al. 1985, p.101), as "the base of the logarithms is an arbitrary choice" (Lance and Williams 1966), and as "if one wanted to confine H always to the range 0 to 1, then logarithms to base m can be used for characters with m states ... this is in effect just the same as multiplying H by a normalizing constant." (Pankhurst 1991, p. 192).

However, the base of the logarithm has practical implications. For the example in table 2, the number of taxa is 3, and there are 4 states. Using $\log_3$ gives H = 0.79 to 2 decimal places and using log4 gives H = 0.63. Base m, the number of states, is not ideal (Pankhurst's (1991) use of m rather than n might perhaps have originated as a typographical error). It is better to normalize using base n, the number of taxa, so that the coefficient remains within the range [0,1] as the number of taxa decreases in the course of an identification, which gives the user a consistent impression of whether a character has significant separating power. Normalizing by the number of states achieves a [0,1] range but removes the benefit of multi-state characters and can give a low rank to completely separable taxa if each taxon has multiple (non-shared) states (Table 6 includes an example of such an anomalous ranking). Subsequent authors may have noticed the problem, but to our knowledge none has implemented the solution; for example, Delta-Intkey uses $\log_2$ throughout.

NUMERIC RANGES

Numeric ranges are treated in most respects like other discrete categories, with the extremely useful exception (Dallwitz et al. 2002) that the user specifies a simple value. For example, if Taxon1 allows leaf length 1–3 cm, a calculation when the database is loaded indicates whether this is distinct from the leaf lengths of other taxa or not. If overlap between taxa occurs, such as if Taxon2 allows leaf length 2–3 cm, then discrete component intervals with open '()' or closed '[]' limits can be calculated, which are henceforth treated separately. When the user enters leaf length 2.16 cm, this needs to match a character state allowed for both Taxon1 and Taxon2. In principle, it is easy to translate numeric ranges into sets, and thence into equivalent characters with non-numeric categorical states. There are some problems with doing this, however. The first problem is a conceptual one: numeric ranges as used in keys impose artificial limits; a character such as length is not naturally categorical, but taxonomists routinely cope with the necessary conversions as they design character states, choosing a total range or a “normal range”, or choosing states that describe a statistical distribution (Jardine and Sibson 1970). Here we follow the approach used by Delta-Intkey, assuming that although ranges might be entered in a more complex format, they come to this component of the polyclave software as simple ranges, e.g., leaf length 2–3 cm.

A second problem is that the computer programming involved in interpreting the specified numeric ranges is non-trivial, but that is also routinely dealt with in polyclave software. A complication here is that overlapping ranges can be divided into subintervals in more than one way; in the example above either [1,2)+[2,3] or [1,2]+(2,3] is possible, as are non-minimal subdivisions such as [1,2)+[2]+(2,3]. The software needs to sort all range endpoints for the character states of a character, then work through the sorted list creating the minimum number of required subranges, and when a choice is possible, consistently assigning closed limits on either the left or right ends of intervals. If the program assigns closed limits at the right ends, then the above example with just two taxa produces limits [1,2]+(2,3], and the entered data 2.16 must match character state (2,3] which is allowed for both Taxon1 and Taxon2.

A third problem is the one we are most interested in here, that, in keeping with the assumptions of information theory, it is not correct to calculate the information content of numeric-range data using intersection of sets (Tables 3 and 4). The assumptions of information theory are that the taxa are equally likely, and that within each taxon the various character states are equally likely (Abramson 1963; Cover and Thomas 2006; Shannon 1948). The size of a numeric range is irrelevant (Table 4). However, unshared discrete states, if shared states are also present, become more important if they are subdivided (Table 3). Consequently, when ranges are compared, it is important to count the number of intervals of overlap and non-overlap (Table 4, Table 5), which is a different calculation from that used in set theory, potentially producing a different character ranking.

EXAMPLES

In a polyclave, normalizing the Best character coefficient to the range [0,1] clarifies whether a character warrants evaluation, which could be helpful if a relatively expensive technique such as microscopy or molecular testing is required. We recommend a coefficient that is the information content normalized by the number of taxa (Table 6). A coefficient of 1.0 signals that specifying that one character will resolve all taxa. The unnormalized value lacks this clarity, and normalizing by the number of character states gives an incorrect ranking of the different characters (e.g., flower color in Table 6).

A stand-alone Bestchar program for categorical data is provided that calculates a variety of coefficients for a single categorical character. The source code is dedicated to the Public Domain, as per http://creativecommons.org/publicdomain/zero/1.0/ Three versions are given, one in the Python language (http://www.python.org/) for clarity in handling lists, one in the R language (http://www.r-project.org/) which is heavily used in biodiversity informatics (Kindt and Coe. 2005; Rossi 2011; Chamberlain and Barve 2012; Kembel 2012; Chamberlain et al. 2013; Hijmans et al. 2013; VanDerWal et al. 2013), and source code for a Java applet and application (http://docs.oracle.com/javase/7/docs). The program source files Bestchar.py, Bestchar.R, and BestChar.java, as well as charinput.txt, a sample input data set corresponding to Table 5 char 2 are accessible at https://github.com/NadiaTalent/Bestchar.

Table 3: If shared states are present, subdividing a character state increases weighting of the character in the information statistic. '/' in a list of character states indicates 'or'.

	Char 1	**Char 2**	**Char 3**	**Char 4**
Taxon 1	a	a/b	a/b/c	a/b/c/d
Taxon 2	b	a	a	a
Discrete sets of states	2	2	2	2
Information (in bits)	1.0	0.25	0.33	0.38

Table 4: The extent of a numeric range is irrelevant to the information content (char 1 and char 2, taxon 5), except in so far as it affects the complexity of how numeric ranges overlap. Overlapping ranges that differ at one extremity (char 3) are readily converted to equivalent categories using the minimum number of subintervals needed to distinguish the character states. Char 3 requires five subintervals [1,2), [2,3), [3,4), [4,5), [5], equivalent to the five categories of char 4. '/' in a list of character states indicates 'or'.

	Char 1	**Char 2**	**Char 3**	**Char 4**
Taxon 1	1–2	1–2	1–5	a/b/c/d/e
Taxon 2	3–4	3–4	2–5	b/c/d/e
Taxon 3	5–6	5–6	3–5	c/d/e
Taxon 4	7–8	7–8	4–5	d/e
Taxon 5	9–10	9–10000	5	e
Intkey's Best Character (observed)	2.32	2.32	0.41	0.41
Information (in bits)	2.32	2.32	0.41	0.41
Normalized to the range [0,1] for 5 taxa	1	1	0.18	0.18

DISCUSSION

As polyclave identification systems become larger, with more varied types of data such as mixed morphological and DNA characters, character ranking might become important and widely used. Although a naïve user might have some difficulty grasping the assumptions on which it is based, and might therefore prefer to ignore the statistic, the experienced botanists and taxonomists whom we have asked generally favor the inclusion of such an automatically calculated feature. The single calculated value indicates to the user who is identifying a specimen whether their effort to evaluate a particular character is likely to yield a significant reduction in the remaining search space of possible identifications. A traditional alternative approach, embodied particularly in single-entry keys, is to use hard-coded expert opinion about which characters are most useful or more reliable. However, whenever serious effort is needed to assess the character states of a specimen, as might occur in forensic and some other applications, pre-coded character rankings may not be the best guide.

Few current taxon-identification systems as yet use a fully multi-entry (polyclave) structure, so the possibilities of character ranking are hardly explored. We are convinced that its potential utility will not be appreciated until it is widely and correctly implemented, until polyclave users have seen it provide a useful hint in difficult situations. If the formula used is clearly linked to the

Table 5: Numeric ranges that overlap are not treated as equivalent to sets; rather, the areas of overlap and non-overlap may form a greater number of categories that separately contribute to the information statistic. Non-overlap at one extremity (Table 4) is a lower-entropy situation than with additional non-overlap (char 1, char 2). Delta-Intkey incorrectly analyzes situations of complex overlap in a way that appears to be consistent with combining non-sequential intervals of non-overlap into a set of intervals, rather than treating the intervals separately, using five sets rather than nine components of char 1, to produce 0.41 rather than 0.47. '/' in a list of character states indicates 'or'; anomalous coefficients appear in bold-face; () and [] indicate range limits, open (omitting the endpoint) and closed (including the endpoint) respectively; ∪ indicates a set union operation on intervals.

	Char 1	**Subintervals of char 1**	**Sets of char 1 subintervals**	**Char 2**
Taxon 1	1–10	[1,2),[2,3),[3,4),[4,5),[5,6],(6,7],(7,8],(8,9],(9,10]	[1,2)∪(9,10],[2,3)∪(8,9],[3,4)∪(7,8],[4,5)∪(6,7],[5,6]	a/b/c/d/e/f/g/h/i
Taxon 2	2–9	[2,3),[3,4),[4,5),[5,6],(6,7],(7,8],(8,9]	[2,3)∪(8,9],[3,4)∪(7,8],[4,5)∪(6,7],[5,6]	b/c/d/e/f/g/h
Taxon 3	3–8	[3,4),[4,5),[5,6],(6,7],(7,8]	[3,4)∪(7,8],[4,5)∪(6,7],[5,6]	c/d/e/f/g
Taxon 4	4–7	[4,5),[5,6],(6,7]	[4,5)∪(6,7],[5,6]	d/e/f
Taxon 5	5–6	[5,6]	[5,6]	e
Intkey's Best Character (observed)	**0.41**			0.47
Information (in bits)	0.47	0.47	**0.41**	0.47
Normalized to the range [0,1] for 5 taxa	0.20	0.20	0.18	0.20

established literature on information theory rather than left undocumented or hidden in a proprietary formula, students may find that literature helpful, which could promote the use of the statistic.

We suspect that the overlap in terminology and the use of the information coefficient in multivariate data exploration may have caused some biologists and software developers to assume that character ranking in polyclaves is a complex topic with many possible solutions, but as we have reviewed above, the aim and the approach are actually quite simple. The characters with the highest ranking are those that divide the remaining taxa into evenly sized small groups. Characters with numeric-range values are difficult to deal with taxonomically, but if coded in such a way that they accurately describe taxa, there is no reason to exclude them from the rank calculation.

We emphasize that ranking characters by their information content in an online polyclave is a distinct problem that has its own special requirements. A related problem is to distinguish taxa and devise a taxonomy, which can involve cluster analysis using similarity or dissimilarity measures, and is a large research focus in many areas apart from biological systematics. Another related problem occurs in software aids for developing single-entry (usually binary) identification keys, but the requirements differ and a separation coefficient is commonly used (Hill

Table 6: If the character-ranking coefficient, normalized by the number of taxa, has a value of 1.0, this means that specifying the character state will resolve all taxa, as with the calyx edge character both before and after the user has specified the state of the life cycle character. This helpful hint to the user is not available from the unnormalized Information (initially 1.58 bits). Using the number of states to normalize is incorrect, not reflecting the relative effectiveness of the characters to resolve the taxa. '/' in a list of character states indicates 'or'; anomalous coefficients appear in bold-face.

	Initial conditions				**After choosing Life cycle=perennial**		
	Calyx edge	**Flower color**	**Life cycle**	**Calyx**	**Calyx edge**	**Flower color**	**Calyx**
Taxon 1	crenate	white	annual	glabrous			
Taxon 2	dentate	orange/pink	perennial	glabrous/ pubescent	dentate	orange/pink	glabrous/ pubescent
Taxon 3	cuspidate	pink/red	perennial	pubescent	cuspidate	pink/red	pubescent
#states	3	4	2	2	2	3	2
#taxa	3	3	3	3	2	2	2
Information (in bits)	1.58	1.25	0.92	0.58	1.0	0.50	0.25
Normalized by #states	1.0	**0.63**	0.92	0.58	1.0	0.32	0.25
Normalized by #taxa	1.0	0.79	0.58	0.37	1.0	0.50	0.25

1974; Pankhurst 1991; Burguiere et al. 2013; ETI Bioinformatics undated); in that situation, look-ahead may be important (Quinlan 1986), and polytomies may be undesirable.

To our knowledge, Delta-Intkey is the only currently available system that uses a coefficient related to information theory, and it is currently being reimplemented (Atlas of Living Australia 2011 onwards), which may allow for improvement. We have suggested that the coefficient should be normalized to the range [0,1], in effect dividing by the number of remaining taxa, to make it more clearly interpretable to the user. We have also discussed the treatment of characters with numeric-range data, which requires special care to remain consistent with information theory.

Computational expense is an issue. The description of Actkey (Brach and Song 2005) suggested that a pre-calculated replication of Intkey's Best Character would be used to sort the characters, and it would not be recalculated in later stages of the identification and would therefore become unreliable. We would argue against taking that approach if at all possible because character ranking could be particularly useful after the most readily available data have been used, when it may be necessary to decide whether to use an expensive test. At late stages like this, the number of remaining taxa is probably reduced, and the calculation therefore becomes more feasible.

ACKNOWLEDGEMENTS

The authors are grateful to Graeme Hirst, Sara Scharf, and Don Tarnawski for helpful discussion, to two anonymous reviews for insightful suggestions, and to the iEvoBio2012 Organizing Committee for agreeing to include NT among their number, an experience that produced a couple of serendipitous insights that encouraged submission of this manuscript.

REFERENCES

Abbott, L. A., F. A. Bisby, and D. J. Rogers. 1985. Taxonomic analysis in biology: Computers, models, and databases. Columbia University Press, New York.

Abramson, N. 1963. Information theory and coding. McGraw-Hill, New York.

Alexander, G. 2006. SLIKS-Alike Interactive Key Software (SAIKS). Accessible at http://www.galexander.org/saiks/README.

Atlas of Living Australia. 2011 onwards. open-delta: A Java port of the DELTA – DEscription Language for TAxonomy suite. Accessible at http://code.google.com/p/open-delta/.

Brach, A. R. and H. Song. 2005. ActKey: a Web-based interactive identification key program. Taxon. 54:1041–1046.

Burguiere, T., F. Causse, V. Ung, and R. Vignes-Lebbe. 2013. IKey+: a new single-access key generation web service. Syst. Biol. 62:157–161.

Chamberlain, S., Boettiger, C., Ram, K. and Barve, V. 2013. Package ‘rgbif’: Interface to the Global Biodiversity Information Facility API methods. Accessible at http://cran.r-project.org/web/packages/rgbif/rgbif.pdf.

Chamberlain, S. and Barve, V. 2012. Package ‘rvertnet’: Search VertNet database from R. Accessible at http://cran.r-project.org/web/packages/rvertnet/rvertnet.pdf.

Christensen, K. I. 1999. MEKA – An introduction to the use of Meacham’s Multiple-Entry Key Algorithm. University of Copenhagen.

Cover, T. M. and J. A. Thomas. 2006. Elements of information theory. John Wiley & Sons, Inc., Hoboken, New Jersey.

Dale, M.B., M. Beatrice, R. Venanzoni, and C. Ferrari. 1986. A comparison of some methods of selecting species in vegetation analysis. Coenoses, 1, 35–52.

Dallwitz, M., T. Paine, and E. Zurcher. 1998/2012. Principles of interactive keys. Accessible at http://delta-intkey.com/www/interactivekeys.htm.

Dallwitz, M. J. 1974. A flexible computer program for generating identification keys. Systematic Zoology. 23:50–57.

Dallwitz, M. J. 1993. Delta and Intkey. Pp. 287–296 in R. Fortuner, ed. Advances in computer methods for systematic biology: Artificial intelligence, databases, computer vision. The Johns Hopkins University Press, Baltimore.

Dallwitz, M. J. 2010. Overview of the DELTA System. Accessible at http://delta-intkey.com/www/overview.htm.

Dallwitz, M. J., T. A. Paine, and E. J. Zurcher. 2002. Interactive identification using the Internet. Pp. 23–33 in H. Saarenmaa, and E. S. Nielsen, eds. Towards a global biological information infrastructure — challenges, opportunities, synergies, and the role of entomology, European Environment Agency Technical Report 70. EEA, Copenhagen.

Dallwitz, M. J., T. A. Paine, and E. J. Zurcher. 2010. User’s guide to the DELTA System: a general system for processing taxonomic descriptions. Accessible at http://delta-intkey.com/www/uguide.htm.

de Lamarck, J. B. P. A. d. M. 1778. Flore françoise; ou, Description succincte de toutes les plantes qui croissent naturellement en France. Disposée selon une nouvelle méthode d’analyse, & à laquelle on a joint la citation de leurs vertus les moins équivoques en médecine, & de leur utilité dans les arts. L'imprimerie Royale, Paris. Accessible at http://www.biodiversitylibrary.org/item/38206.

Duke, J. A. 1969. On tropical tree seedlings I. Seeds, seedlings, systems, and systematics. Annals of the Missouri Botanical Garden. 56:125–161.

Duncan, T. and C. A. Meacham. 1986. Multiple-entry keys for the identification of Angiosperm families using a microcomputer. Taxon. 35:492–494.

Duncan, T. and C. A. Meacham. 1987. Meka manual. University Herbarium, University of California, Berkeley, California, USA.

ETI Bioinformatics. undated. Linnaeus II. Accessible at http://www.eti.uva.nl/products/linnaeus.php.

Gower, J. C. 1975. Relating classification to identification. Pp. 251–263 in R. J. Pankhurst, ed. Biological identification with computers. Academic Press, London and Orlando.

Guala, G. F. 2004–12. SLIKS: Stinger’s Lightweight Interactive Key Software. Accessible at http://www.stingersplace.com/SLIKS/.

Hagedorn, G., G. Rambold, and S. Martellos. 2010. Types of identification keys. Pp. 59–64 in P. L. Nimis, and R. V. Lebbe, eds. Tools for Identifying Biodiversity: Progress and Problems. EUT Edizioni Università di Trieste, Trieste.

Hijmans, R.J., Phillips, S., Leathwick, J. and Elith, J. 2013. Package 'Dismo': Species distribution modeling. Accessible at http://cran.r-project.org/web/packages/dismo/dismo.pdf.

Hill, L. R. 1974. Theoretical aspects of numerical identification. Int. J. Syst. Bacteriol. 24:494–499.

Jardine, N. and R. Sibson. 1970. Quantitative attributes in taxonomic descriptions. Taxon 19:862–870.

Kembel, S. 2012. Biodiversity analysis in R: CSEE R Workshop 2012. Accessible at http://phylodiversity.net/skembel/r-workshop/biodivR/SK_Biodiversity_R.html.

Kindt, R. and R. Coe. 2005. Tree diversity analysis: A manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre, Nairobi, Kenya. Accessible at http://www.worldagroforestry.org/downloads/publications/PDFs/B13695.pdf.

Lance, G. N. and W. T. Williams. 1966. Computer programs for hierarchical polythetic classification (“similarity analyses”). The Computer Journal. 9:60–64.

Lobanov, A. L., W. F. Schilow, and L. M. Nikritin. 1981. Zur Anwendung von Computern für die Determination in der Entomologie. Dtsch. Entomol. Z. 28:29–43.

Lucidecentral.org. 2010. About Lucid. Accessible at http://www.lucidcentral.org/Home/AboutLucid/tabid/203/language/en-US/Default.aspx.

MacDonald, D. K. C. 1952. Information theory and its application to taxonomy. Journal of Applied Physics. 23:529–531.

MacLeod, N., ed. 2007. Automated taxon identification in systematics: Theory, approaches and applications. CRC Press, Taylor and Francis Group, Boca Raton.

Martellos, S. 2010. Multi-authored interactive identification keys: The FRIDA (FRiendly IDentificAtion) package. Taxon. 59:922–929.

Morse, L. E. 1971. Specimen identification and key construction with time-sharing computers. Taxon. 20:269–282.

Morse, L. E. 1975. Recent advances in the theory and practice of biological specimen identification. Pp. 11–52 in R. J. Pankhurst, ed. Biological identification with computers. Academic Press, London and Orlando.

Osborne, D. V. 1963. Some aspects of the theory of dichotomous keys. New Phytol. 62:144–160.

Pankhurst, R. J. 1975. Identification by matching. Pp. 79–91 in R. J. Pankhurst, ed. Biological identification with computers. Academic Press, London and Orlando.

Pankhurst, R. J. 1978. Biological Identification: The Principles and Practice of Identification Methods in Biology. Edward Arnold, London.

Pankhurst, R. J. 1991. Practical taxonomic computing. Cambridge University Press, Cambridge.

Pankhurst, R. J. and R. R. Aitchison. 1975. A computer program to construct polyclaves. Pp. 73–78 in R. J. Pankhurst, ed. Biological identification with computers. Academic Press, London and Orlando.

Payne, R. W. and D. A. Preece. 1980. Identification keys and diagnostic tables: a review. J. Roy. Stat. Soc. Ser. A. (Stat. Soc.). 143:253–292.

Podani, J. 2000. Introduction to the exploration of multivariate biological data. Backhuys Publishers, Leiden.

Quinlan, J. R. 1986. Induction of decision trees. Machine Learning. 1:81-106.

Rosatti, T. J. undated. Electronic, interactive identification keys for California plants Using MEKA (Multiple-Entry Key Algorithm). Accessible at http://ucjeps.berkeley.edu/keys/.

Rossi, J.-P. 2011. rich: An R Package to Analyse Species Richness. Diversity. 3:112–120.

Shannon, C. E. 1948. A mathematical theory of communication. The Bell System Technical Journal. 27:379–423, 623–656.

Shwayder, K. 1971. Conversion of limited-entry decision tables to computer programs — a proposed modification of Pollack’s algorithm. Communications of the ACM. 14:69–73.

Shwayder, K. 1974. Extending the information theory approach to converting limited-entry decision tables to computer programs. Communications of the ACM. 17:532–537.

Sneath, P. H. A. and R. R. Sokal. 1973. Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman and Company, San Francisco.

Spooner, A. and A. Chapman. 2007. DELTA Intkey Tutorial. Western Australian Herbarium. Accessible at http://florabase.dec.wa.gov.au/help/keys/intkey_tutorial.pdf.

Toda, M. J., K. Matsushita, and S. F. Mawatari. 2004. Biological Classification and Identification System (BioCIS). Neo-Science of Natural History: Proceedings of International Symposium on "Dawn of a New Natural History — Integration of Geoscience and Biodiversity Studies", Sapporo. Accessible at http://hdl.handle.net/2115/38489.

Ung, V., G. Dubus, R. Zaragüeta-Bagils, and R. Vignes-Lebbe. 2010. Xper²: introducing e-Taxonomy. Bioinformatics 26:703–704.

VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L. and Storlie, C. 2013. Package 'SDMTools': Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. Accessible at http://cran.r-project.org/web/packages/SDMTools/SDMTools.pdf.

Voss, E. G. 1952. The history of keys and phylogenetic trees in systematic biology. Journal of the Scientific Laboratories, Denison University 43:1–25.

Walter, D.E. and S. Winterton, S. 2006. Keys and the Crisis in Taxonomy: Extinction or Reinvention? Ann. Rev. Entomol. 52:193-208.

Walters, S. M. 1975. Traditional methods of biological identification. Pp. 3–8 in R. J. Pankhurst, ed. Biological identification with computers. Academic Press, London and Orlando.