The Theoretical Costs of DNA Barcoding

Monika Piotrowska

Department of Philosophy
University of Utah
Salt Lake City, UT, USA
monika.piotrowska@utah.edu

When a child's piggy bank is full, she can go to a grocery store and pour it into a change-sorting machine. Such a machine will sort all the quarters into one slot, all the dimes into another, and so on, until all the change is sorted and the machine can add up the sums of each slot. Now imagine a similar machine, but this time it sorts insects instead of change. A child who has been collecting bugs has brought them in a jar and pours them into an insect-sorting machine. The machine takes a blood sample of each bug and sorts the samples according to the species to which they belong—identified via a sequence of their DNA. The child is then given a printout of all the information regarding each species whose member she has captured.

Although insect-sorting machines are fictitious, a diagnostic technique that makes the use of short DNA sequences to identify species is real—the technique is known as *DNA barcoding*. Its advocates claim that DNA barcodes can be used to sort specimens into species at a rate quicker than that of practicing taxonomists and the resulting groupings are usually the same. A number of studies purport to have confirmed this; they are used to justify the (sole) use of barcodes for species delimitation, as they demonstrate that groupings based on barcodes match those of an existing taxonomy. My aim in this article is to call into question just what these studies license and whether the sole use of DNA barcodes to delimit species is indeed justified.

Since the debate over DNA barcoding has been a heated one (e.g., Hebert and Gregory 2005; Kipling et al. 2005; DeSalle 2006; Rubinoff 2006), I begin by describing the benefits and limits of DNA barcoding as presented by its advocates not its critics. Next, I argue that due to the mutually dependent relationship between defining and delimiting species, all systems of classification are grounded in theory, even if only implicitly. I then proceed to evaluate DNA barcoding in that context. In particular, I focus on the barcoders' use of a sharp boundary by which to delimit species, arguing that this boundary brings along additional theoretical commitments inconsistent with the way taxonomists conceive of species, viz., as entities that have vague boundaries and that cannot be defined by any single attribute other than ancestry. Given these inconsistencies, I conclude that even if groupings based on DNA barcodes match those of an existing taxonomy, the two systems of classification are not necessarily tracking the same entities, i.e., species.

What Is DNA Barcoding?

DNA barcoding uses a 648-bp region of the mitochondrial genome—*sytochrome c oxidase 1* (CO1)—to identify organisms as belonging to a particular species (Hebert et al. 2003). In 2003, the founder of DNA barcoding, Paul Hebert, proposed the compilation of a public library of DNA barcodes that would be linked to species already defined by taxonomists. Currently, the database contains 37,639 unique barcodes that match previously described species (Ratnasingham and Hebert 2007). The point of the barcode library is to provide a new master key for identifying species, and hence to make it possible for an unknown specimen to be identified as a member of a known species if its sequence closely matches one in the database. Thus, for instance, the diversity of ant species in Madagascar was assessed rapidly by running the collected barcodes against the library reference code (Smith et al. 2005). Conversely, if the CO1 sequence of an unknown specimen does not match any of the barcodes in the database, it can suggest the existence of a new species: "Newly encountered species will ordinarily signal their presence by their genetic divergence from known members of the assemblage" (Hebert et al. 2003: 318). Hence, when the DNA barcodes of 260 North American bird species were examined, four of the varieties of known bird species came out as potential new species because their sequences diverged past the established sequence similarity threshold (Hebert et al. 2004).

Initially, the idea of using DNA barcodes to discover new species was heavily criticized by taxonomists (e.g., Lipscomb et al. 2003; Mallet and Willmott 2003; Seberg et al. 2003). However, DNA barcoding advocates have since restated their position clarifying that barcodes, by themselves, are never sufficient to *describe* new species—only to *identify* members of known species:

There is an important distinction between "describing" and "delimiting" species, but a conflation of the two has created uneasiness about the use of DNA barcodes as the foundation of future taxonomic descriptions. We emphasize that DNA barcoding seeks merely to aid in *delimiting* species—to highlight genetically distinct groups exhibiting levels of sequence divergence suggestive of species status. By contrast, DNA barcodes—by themselves—are never sufficient to describe new species. (Hebert and Gregory 2005: 853)

Hence, the primary use of DNA barcodes is to match CO1 sequences to "known" species, i.e., species that have already been characterized by taxonomists. According to advocates, traditional methodologies must still be used for species discovery. Returning to the North American bird study example, in order to determine whether the four unmatched specimens were in fact members of new species, they would need to undergo further taxonomic analysis (involving multiple genes, ecological characters, morphological characters, behavior, population biology, geography, etc.).

The claim, then, is that DNA barcodes, by themselves, are not sufficient to describe new species but that they suffice to *delimit* known species. How so? How can members of known species be identified solely with a "scan" of their barcode? The reason barcodes can discriminate between species is due to the rapid sequence change in mitochondrial DNA (mtDNA), which allows for the accumulation of differences between populations that have only been separated for brief periods of time (Hebert et al. 2004). Thus, mtDNA sequence divergence between species is expected to be much larger than within species.

At least this was the initial hypothesis. A number of studies have since been conducted in an effort to find a correspondence between the categories of species established by taxonomists and those inferred by DNA barcoding proponents. These studies were aimed at confirming the presence of a unique barcode in every species and identifying the sequence similarity threshold particular to each. However, what the studies have shown is that certain kinds of organisms do not have unique barcodes at the CO1 location; examples include amphibians (Vences et al. 2005), plants (Kress et al. 2005), sea anemones, corals, and some jellyfish (Hebert et al. 2003). As for the organisms that do have unique barcodes, the sequence similarity criterion for species membership was set to no more than 3% divergence for all insects, but 2% for birds and mammals (Hebert et al. 2003).

Adjustments were thus made to the initial DNA barcode hypothesis as it was tested against the established taxonomic groups. Recent claims put forth by Hebert hold that CO1 sequences are 99.75% identical between members of one species and less than 97.5% identical across members of various species (Strauss 2006). The small margin between being included and excluded from a species is likely what accounts for the high success rate of DNA barcoding, where "success" means the proportion of specimens identified with the use of DNA barcodes that match the species distinguished by prior taxonomic work. So far, DNA barcoding studies that looked at species of birds, spiders, fish, springtails, and several arrays of Lepidoptera have been close to 100% accurate in their discrimination (Hebert et al. 2004; Hogg and Hebert 2004; Barrett and Hebert 2005; Ward et al. 2005; Hajibabaei et al. 2006). These and other studies are held up as strong evidence to justify the utility of DNA barcodes as good diagnostic characters for species delimitation.

What are the Theoretical Commitments of DNA Barcoding?

If these studies license the sole use of barcodes for species delimitation, then they do so under the following assumption: If members of a species share a unique similarity, then that similarity may be used, by itself, to pick out members of a species.[1] But if barcoders want to use the CO1 sequence to pick out members of a species, they need a *theory* about species that explains why the barcode is the *relevant* similarity. Otherwise, their categories would seem to be merely based on CO1 sequence similarity, and categories based on similarity without theory run the risk of being circular. Let me explain.

Organisms can be similar with respect to different properties, which means that any two organisms similar from one point of view may be dissimilar from another. As Nelson Goodman put it, "Any two things have exactly as many properties in common as any other two" (Goodman 1972: 443). This is because we could classify things according to the category of being "less than 100 meters long," being "less than 101 meters long," etc., leading to an infinite number of similarities and dissimilarities between any two organisms. Consequently, proponents of DNA barcoding need to have an argument for why the barcode is the *right* similarity. In other words, why should we group organisms according to their CO1 sequence and not by some other property? If the only theory of species on the table is one that invokes a CO1 sequence similarity criterion, then DNA barcoding advocates do not have a good answer to that question. And if two objects, or two organisms, are similar only because they are in the same category (e.g., if two animals from the category *Homo sapiens* are similar only because they have the *Homo sapiens* CO1 sequence), then any account of categorization based on similarity is circular.

In spite of the looming circularity objection, DNA barcoders have not articulated their own theory of species, probably because they take themselves to be merely *delimiting* species, whereas the theoretical task of *defining* species is delegated to taxonomists (Hebert and Gregory 2005). Accordingly, the taxonomists are the ones who deal with the hard questions, such as "What are species?," "What constitutes a speciation event?," etc. DNA barcoders are only concerned with group membership, which is a relatively simple "yes" or "no" identification procedure based on *predefined* units (Rubinoff 2006).

Barcoders are right to point out the distinction between defining species and deciding how these entities are to be delimited (Sites and Marshall 2003), but they fail to mention the intimate connection between the two. Which individuals are allocated to a given species depends on the criteria used to delimit species, and the delimiting criteria are determined by the concept of what a species is. This is why different concepts lead to incongruent species boundaries. Thus, a recent literature survey revealed a 48.7% higher count of species with the application of a *phylogenetic* species concept compared with studies applying the *biological* species concept to the same set of organisms (Agapow et al. 2004). Similarly, the allocation of members has a direct influence on which segments of separately evolving lineages end up corresponding to the category "species." The point is that different species concepts can lead to different members of species, and different members can lead to different species concepts.

The surplus of species concepts and membership criteria is in large part due to the complex way in which species evolve. While the formation of species as parts of evolving lineages can be explained in terms of a few general evolutionary processes (mutation, natural selection, migration, and genetic drift), the characters affected by these processes are highly diverse. They may be genotypic or phenotypic; qualitative or quantitative; selectively advantageous, disadvantageous, or neutral. And they may involve many different aspects of organismal biology, including genetics, development, morphology, physiology, and behavior (De Queiroz 2007: 881). Species may acquire any number of the above characters as they separate and diverge from one another, but the evidence is usually indirect (Vogler and Monaghan 2007: 5). This is the reason why delimiting criteria vary and are ultimately determined by an underlying species concept.

When DNA barcoders rely on the CO1 sequence to allocate unidentified specimens to particular species, the allocation has a direct influence on which segments of the separately evolving lineages end up corresponding to the category "species." In particular, the sharp boundary by which DNA barcoders delimit species reveals their theoretical commitment as to what constitutes a speciation event. Here is why. If a specimen were found with a mutation at the CO1 location, which changed similarity to below 97.5%, we already know that DNA

barcoders would not feel confident assigning the diverged sequence to a different species—especially if there was no match for it in the barcode library. (This goes back to one of their commitments—that barcodes, by themselves, are never sufficient to describe new species.) What barcoders would do, however, is exclude the specimen from the current species and put it aside for traditional taxonomists to make the final analysis—which means that they are willing to toss the specimen into the "unknown" species pile. But in doing so they actually make a claim about what is required for a speciation event to occur, because by throwing the diverged sequence into the "unknown" pile they are claiming that a specimen with a sequence less than 97.5% similar to the sequences of other members is itself not a member. Thus, the 97.5% mark is their cutoff point for either belonging to or being excluded from a species, which is why the specimen with a sequence that drops below the cutoff point will no longer belong to that species but to some other (whether or not this other species has yet been identified is irrelevant). Consequently, DNA barcoders must be committed to the idea that a mutation at the CO1 location that changes similarity to below 97.5% constitutes a speciation event.

Another way of thinking about this theoretical commitment is inspired by the perspective of conservation biology: DNA barcoders are committed to a certain view regarding the percent divergence at the CO1 location that needs to be in place before diversity assessments (declaring the number of separate species present) can be made. Either way, the picture that emerges is that for barcoders the way to pick out species from separately evolving lineages is by looking for CO1 sequences that have diverged past the 97.5% mark.

Are DNA Barcoders and Taxonomists Tracking the Same Entities?

There are at least two problems with this characterization of species. First, setting a sharp boundary between species (based on sequence divergence levels, e.g., >3%) conflicts with the kind of continuity found in nature. Species change continuously and are gradually transformed into novel ones. The process of speciation does not include sudden discontinuities that could be used as a specific boundary (Cain 1954: 107). The apparent state of equilibrium of species is just a current representation of a continuously evolving lineage, a mere illusion fostered by the shortness of the human lifespan (Gross 1988: 226). In all but a few cases, speciation is a long and gradual process such that there is no principled way to draw a precise boundary between one species and the next. The boundaries between species are vague (Hull 1965). Given this vagueness, asking for a sharp boundary by which to delimit species is an implausible demand. It is as implausible as demanding that a precise number of dollars marks the boundary between rich and poor or that a precise number of hair marks the boundary

between bald and not bald. If being a species along with being a heap of stones, being bald, and being rich are all concepts beset by line-drawing problems (Sober 1980: 356), then the DNA barcode is an arbitrary cutoff for species delimitation.

The second problem with the barcoders' characterization of species is their emphasis on the degree of similarity over common ancestry. Species are segments of separately evolving lineages; they are properly defined by the only attribute they possess that cannot change during the course of evolution— their ancestry. How does one determine which organisms are parts of which lineage? Genetic similarities may provide the evidence, since they are typically inherited along ancestral lines, but the determining factor is the relation among organisms. The same rule applies to family trees. Being part of my immediate family turns on my siblings, my parents, and I having certain biological relations to one another, not on our having similar features. I may have inherited my mother's gene for cystic fibrosis and my father's gene for tongue rolling, but the genetic similarities between us are not the reason we are part of the same family; we are part of the same family because we are appropriately causally connected (Hull 1976, 1978; see also Ereshefsky 2007).

DNA barcoders have denied ancestry a central role in taxonomy by using only one line of evidence to delimit species, viz., sequence similarity at the CO1 location. Consequently, they have fallen victim to circular reasoning: thinking that a shared similarity between the members of a species is sufficient for species membership. By not considering another line of evidence to test their hypothesis, they placed themselves in a position from which they cannot break out of that circle. Traditional taxonomists avoid this problem (and thus preserve the scientific aspect of species delimitation) by relying on multiple genes, ecological characters, morphological characters, behavior, population biology, and geography when identifying species (DeSalle et al. 2005). DNA barcoders, however, have favored a single attribute (other than ancestry), and as a result, they have reversed the priority of the evidence with the priority of the natural process, giving rise to the evidence.

In sum, the barcoders' use of a sharp boundary by which to delimit species amounts to smuggling in theoretical claims as to what constitutes a speciation event—claims incompatible with the way taxonomists conceive of species. Given these inconsistencies, I conclude that even if the groupings based on DNA barcodes match those of an existing taxonomy, such findings are not sufficient to prove that the two systems of classification are in fact tracking the same entities.

Acknowledgments

Note

1. The reason the unique similarity has to be used by itself to pick out members of a species is that the barcode was never meant to be one among many diagnostic characters. It was meant as *the only one*. If it were otherwise, DNA barcoding would be nothing new; the use of DNA for the identification of species—along with other diagnostic techniques—goes back to the beginning of molecular systematics (Kipling et al. 2005: 844).

References

Agapow P, Bininda-Emonds ORP, Crandall KA, Gittleman JL, Mace GM, Marshall JC, Purvis A (2004) The impact of the species concept on biodiversity studies. Quarterly Review of Biology 79: 161–179.

Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. Canadian Journal of Zoology 83: 481–491.

Cain AJ (1954) Animal Species and Their Evolution. London: Hutchinson.

De Queiroz K (2007) Species concepts and species delimitation. Systematic Biology 56: 879–886.

DeSalle R (2006) Species discovery versus species identification in DNA barcoding efforts: Response to Rubinoff. Conservation Biology 20: 1545–1547.

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: Taxonomy, species delimitation, and DNA barcoding. Philosophical Transactions of the Royal Society of London B 360: 1905–1916.

Ereshefsky M (2007) Species. In: Stanford Encyclopedia of Philosophy (Zalta EN, ed). Available at: http://plato.stanford.edu/archives/fall2008/entries/species/

Goodman N (1972) Problems and Projects. New York: Bobbs-Merrill.

Gross AG (1988) Philosophy versus science: The species debate and the practice of taxonomy. In: PSA 1988 (Fine A, Forbes M eds), 223–230. East Lansing, MI: Philosophy of Science Association.

Hajibabaei M, Janzen DH, Burns JN, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proceedings of the National Academy of Sciences USA 103: 968–971.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321.

Hebert PDN, Gregory RT (2005) The promise of DNA barcoding for taxonomy. Systematic Biology 54: 852–859.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. PLoS Biology 2(10): e312. Available at: http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0020312

Hogg ID, Hebert PDN (2004) Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. Canadian Journal of Zoology 82: 749–754.

Hull D (1965) The effect of essentialism on taxonomy—two thousand years of stasis. British Journal for the Philosophy of Science 15: 314–326; 16: 1–8.

Hull D (1976) Are species really individuals? Systematic Zoology 25: 174–191.

Hull D (1978) A matter of individuality. Philosophy of Science 45: 335–360.

Kipling WW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 54: 844–851.

Kress JW, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences USA 102: 8369–8374.

Lipscomb D, Platnick N, Wheeler Q (2003) The intellectual content of taxonomy: A comment on DNA taxonomy. Trends in Ecology and Evolution 18: 65–66.

Mallet J, Willmott K (2003) Taxonomy: Renaissance or tower of Babel? Trends in Ecology and Evolution 18: 57–59.

Ratnasingham S, Hebert PDN (2007) BOLD: The barcode of life data system. Molecular Ecology Notes 7: 355–364.

Rubinoff D (2006) Utility of mitochondrial DNA barcodes in species conservation. Conservation Biology 20: 1026–1033.

Seberg O, Humphries CJ, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends in Ecology and Evolution 18: 63–65.

Sites JW, Marshall JC Jr (2003) Delimiting species: A renaissance issue in systematic biology. Trends in Ecology and Evolution 18: 462–470.

Smith AM, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: The ants of Madagascar. Philosophical Transactions of the Royal Society of London B 360: 1825–1834.

Sober E (1980) Evolution, population thinking, and essentialism. Philosophy of Science 47: 350–383.

Strauss S (2006) The barcode of life takes flight. University Affairs (March 13). Available at: http://www.universityaffairs.ca/the-barcode-of-life-takes-flight.aspx

Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: Chances and challenges. Philosophical Transactions of the Royal Society of London B 360: 1859–1868.

Vogler AP, Monaghan MT (2007) Recent advances in DNA taxonomy. Journal of Zoological Systematics and Evolutionary Research 45: 1–10.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society of London B 360: 1847–1857.