University of Malaya

From the SelectedWorks of John James Wilson

2011

Taxonomy and DNA Sequence Databases: A Perfect Match?

John James Wilson, University of Malaya

Available at: https://works.bepress.com/johnjames_wilson/3/
Taxonomy and DNA sequence databases: A perfect match?

John James Wilson

Department of Integrative Biology, University of Guelph, Guelph, Ontario, N1G 2W1, Canada
e-mail: wilso04@gmail.com
Received: 3 May 2011; accepted: 15 June 2011

Summary
Despite the declining number of traditional taxonomists, our knowledge of Earth’s biodiversity continues to grow in the form of DNA sequence data. Freely available through online databases, analyses of sequence datasets are increasingly used as an alternative for the traditional taxonomic process. Species identifications have become “DNA barcoding,” new species discoveries are characterised by genetic divergences, and traditional classification has been supplanted by molecular phylogenetics. These developments are illustrated through a case study investigating the identities of *Taygetis* butterflies of Costa Rica. Here I review prospects and problems with the molecularization of taxonomy and the key role of publicly available nucleotide sequence databases in efforts to catalogue diversity of life.

Keywords
Taxonomy; systematics; DNA barcoding; species identification; species description; phylogenetics; classification; Lepidoptera; *Taygetis*

Introduction
The wealth of biological data freely available online through the International Nucleotide Sequence Database Collaboration (INSDC) comprising GenBank, EMBL (European Molecular Biology Laboratory) and DDBJ (DNA Data Bank of Japan), represents an unparalleled resource to the field of taxonomy and efforts to catalogue biological diversity. In this article I discuss the advantages and perils associated with three main applications of publicly available DNA sequences in taxonomy: (i) species identifications, (ii) facilitating species discoveries, (iii) a source of characters and/or taxa for phylogenetic reconstructions, and discuss the crucial role they could play in a DNA taxonomy of life. These developments are illustrated through a case study investigating the identities of the *Taygetis* butterflies of Costa Rica.
Species identifications

The “taxonomic impediment” and general decline of the traditional taxonomic workforce (Gaston and O’Neill, 2004) has resulted in biologists looking to new and inventive methods to identify specimens, whether, for example, for medical, ecological or agricultural purposes (e.g., Weeks et al., 1999; Hebert et al., 2003a). Advances in high throughput DNA sequencing (Shendure et al., 2004) and reductions in costs (Hajibabaei et al., 2005), means BLAST sequence similarity searching (BLAST is an algorithm used to search DNA sequence databases for a “best hit”; Wheeler et al., 2003) and the GenBank database are frequently employed as an identification tool, to determine the species or higher taxonomic grouping of an unknown specimen or sample from which a DNA sequence can be obtained. The use of short DNA sequences for species identification, DNA barcoding (Hebert et al., 2003a), with careful consideration of the gene region used, has been shown to be very effective in a wide variety of taxonomic groups (Hebert et al., 2003b; Waugh, 2007; Floyd et al., 2009). Sequence-based identification could enable species diagnosis from any life history stage and parts of whole organisms (Savolainen et al., 2005). This includes larval stages difficult to identify with traditional methods (Janzen et al., 2005), social insects in which several casts have “unrelated” morphologies (Smith et al., 2005) and historical fragments lacking species-specific morphological features (Noonan et al., 2005). Novel applications have included parasite forensics (De Bruyne et al., 2005), the fight against trade in endangered species (Hsieh et al., 2003), and classification of snake venoms (Pook and McEwing, 2005). It is not always necessary to sequence a DNA fragment from the unknown sample in order to use sequence databases as a tool in species identification. Sequences from GenBank have been used to develop cheaper diagnostic tools such as PCR assays (e.g. Chapman et al., 2003).

Exploitation of species barcodes, analogous with retail barcodes, first requires the assembly of a comprehensive database that links organisms and their sequences (Savolainen et al., 2005). Presently databases have very uneven distributions of sequences among taxa. Intensely studied groups and model organisms (e.g., Arabidopsis thaliana, Drosophila melanogaster) have many sequences and even entire genomes available. A few genes have been sequenced for many taxa but the vast majority of species diversity is unrepresented (Sanderson et al., 2003). Even amongst mammals diversity is insufficiently covered; ancient rabbit fragments could not be confidently identified due to lack of reliable conspecific reference sequences (Yang et al., 2005). This raises the possibility that inappropriate reference sequences could be applied resulting in spurious species diagnoses (Baker et al., 1996). Best BLAST hit, the simplest method of taxonomic assignment, is “essentially useless” when no relatives have appropriate sequences in GenBank (Tringe and Rubin, 2005). Tautz et al. (2003) suggest a DNA sequence is provided alongside all future taxonomic samples and species descriptions, and the current barcoding initiatives (www.ibol.org) go a long way to bridge the gap (for some major eukaryote groups at least). DNA barcoding currently operates through a system of “reciprocal illumination”: improvement of the DNA sequence database through classical taxonomy, but also improvement of classical taxonomy through
flagging potential cryptic diversity and other taxonomic issues via DNA sequence analysis (e.g. Wilson et al., 2010).

Some authors argue GenBank is unsuitable for taxonomic purposes due to lack of provision to include morphological, biogeographical, and ecological information associated with the sequence entry (Tautz et al., 2003). However, the concept of “type sequences” with voucher specimens authenticated by experts on the taxa and with associated taxonomic data is becoming reality. In 2004 NCBI (National Centre for Biotechnology Information), GenBank’s operators, sealed a partnership with the Consortium for the Barcode Of Life whereby “barcode standard” DNA sequences with relevant supporting data can now be archived with the INSDC (Figure 1; Hanner, 2005; Savolainen et al., 2005) with the keyword “BARCODE” attached. The BOLD (Barcode of Life Data Systems; www.barcodinglife.org; Ratnasingham and Hebert, 2007) is fast approaching 100,000 formally described species with barcodes (cytochrome c oxidase subunit 1 - COI - sequences), most associated with images, specimen and collection data, presenting an unparalleled opportunity for rapid, accurate species identification.

Figure 1. Possible route of species identification using the GenBank database and the criteria for assigning “BARCODE” status in INSDC (Achalarus albociliatus data from Hajibabaei et al., 2006).
Species discoveries

Some taxonomists propose a wider use of DNA sequences and see them as facilitating species discoveries based on cluster or phylogenetic analyses of homologous sequences (Monaghan et al., 2005; Savolainen et al., 2005; Pons et al., 2006). Despite a history of cryptic species discoveries within morphological indiscriminate groups via DNA evidence, this application appears much more controversial than species identification (Will et al., 2005; DeSalle, 2006; Nielsen and Matz, 2006). However, many published empirical “barcoding” study have highlighted potential new species, with species delination at least partially relying on DNA data (e.g., Hebert et al., 2004; Monaghan et al., 2005; Hajibabaei et al., 2006).

For routine specimen identification the BLAST algorithm is currently the only way to search GenBank for “matching” sequences, however, no one would seriously present lack of similar GenBank sequences as conclusive proof of a new undescribed species (Nielsen and Matz, 2006). The exception would be uncultured microbes where sequence similarity within rRNA genes is the accepted standard method of defining species (Venter et al., 2004). Neighbour-joining trees have frequently been used in barcoding analysis of COI sequence datasets; justified by barcoders as providing necessary speed of analysis for large datasets typical of barcoding studies, and because “the goal of barcoding is to provide species identification based on sequence similarity rather than to reconstruct deeper phylogenetic relationships accurately” (Ball et al., 2005). The observation of large inter- and low intra-species sequence variation, the “barcoding gap,” promises easy identification of pre-existing species (Hebert et al., 2003a; Hebert et al. 2010; but see Wiemers and Fiedler, 2007).

However, there are both theoretical and practical concerns regarding phenetic threshold values (DeSalle et al., 2005; Meyer and Paulay, 2005; Monaghan et al., 2005), and this becomes an even more controversial practice when they are used to delimit new species. When analyzing COI sequence datasets, 3% divergence (Kimura 2-parameter) has been cited as sufficient genetic distance to characterize different species (Hebert et al., 2003b). Alternatively, divergences greater than a minimum threshold of one tenth of the average p-distance found between well established species in a lineage have been interpreted as indication of the possibility of cryptic species or misdiagnosis of specimens (Hebert et al., 2004; Monaghan et al., 2005; Smith et al., 2006). Meyer and Paulay (2005) determined that for three mollusc datasets a 3% threshold minimizes over-splitting but results in significant over-lumping of species (termed evolutionary significant units). It seems inevitable that taxonomists using phenetic thresholds will have to constantly revise similarity cut-offs from group to group making the delineation of species using distances fairly subjective (DeSalle et al., 2005; Meyer and Paulay, 2005).

Arguably, taxonomists are more willing to accept an evolutionary approach to new species recognition with DNA sequence data. This would be based on the principles of phylogenetic systematics, where evolutionary entities are inferred with shared derived characters (synapomorphies). Suitable methods of quantitative species delimitation might be to search for diagnostic sequence variation, then determine species limits
based on quantitative methods (see Sites and Marshall, 2003; Monaghan et al., 2005). One possibility is statistical parsimony haplotype network analysis (Templeton et al., 1992; Hart and Sunday, 2007). This method subdivides variation in a sequence dataset based on the level of homoplasy within the data themselves providing a relative measure of divergence rather than a pre-determined phenetic cut-off value (Monaghan et al., 2006). Other alternatives being investigated include elegant methodologies incorporating principles of population genetics (Nielsen and Matz, 2006) and an interpretation of branch length as the species level boundary (Monaghan et al., 2005; Pons et al., 2006; Vuataz et al., 2011). The “barcoder problem” (i.e. which decisionary system to employ to delimit species on the basis of DNA barcode sequences; DeSalle et al., 2005) remains an active area of research and discussion (e.g. Lohse, 2009; O’Meara, 2010).

The problem of species delineation is directly related to the “problem” of species concepts, themselves of perennial interest to biologists due to the unique position of species in the taxonomic hierarchy as the only “real” grouping of natural populations. If identification and discovery of “biological” species is the goal of taxonomy, using phenetic DNA distances to determine species boundaries could suffer the same failings inherent with the partial reality that morphological species equate with biological species, implicit in the traditional methods. Selection of appropriate markers and analytical tools to delimit clusters of interbreeding individuals will remain a challenging problem (Savolainen et al., 2005). Modern evolutionary species concepts with emphasis on monophyly or diagnosability (Baum and Donoghue, 1995) may be more compatible with DNA-based species delineation.

Characters and/or taxa for phylogenetic reconstructions

Though many systematists promote an “integrated” taxonomy and a “total evidence” (i.e., a combination of molecular and morphological characters) approach in phylogenetics, molecular data has distinct advantages over morphological characters in phylogenetic reconstructions (see Scotland et al., 2003) and researchers often obtain sequences from databases to supplement their own sequencing work or increase taxonomic coverage. The large numbers of characters available with sequence data increases robustness and support values of the phylogenetic hypothesis and the nature of DNA evolution allows explicit models of evolution to be incorporated into tree building algorithms (Holder and Lewis, 2003). Huge data matrices comprising unambiguous character states and homology assessments are possible with sequence data (Scotland et al., 2003) and phylogenetic hypotheses derived from molecular data can be used with calibrated molecular clocks to estimate divergence timings of taxa (Savolainen and Chase, 2003). However, there are still long recognised problems such as multiple sequence alignment in length variable regions and the phylogenetic treatment of gaps (Wheeler, 1996).

Short DNA sequences available from databases and their ever-increasing taxonomic coverage could become an unprecedented resource for phylogenetics in addition to being a diagnostic taxonomic tool (Savolainen et al., 2005). It would seem
short-sighted to disregard the potential value of short DNA sequences in phylogenetic analysis especially if they are standardized and used in combination to provide multiple independent sources of evidence, for example, both mitochondrial and nuclear loci. Despite concern over the use of DNA barcodes (specifically COI) in phylogenetics (Wilson, 2010), recent studies have demonstrated the value of COI in resolving deep divergences and report limited incongruence between mitochondrial and nuclear regions (Sihvonen et al. 2011). Dense taxon sampling of databases can limit the problem of missing taxa, a serious concern to accurate recovery of phylogenetic patterns (Wheeler, 2004), as long branches would be broken up in many cases and globally convergent characters become local homologies. With near complete taxon sampling, even short DNA sequences could theoretically resolve even deep phylogenies (see Pollock et al., 2002).

The most informative phylogenetic dataset will include as many taxa and nucleotides as possible while simultaneously limiting the amount of missing data (Yan et al., 2005). An investigation by Sanderson et al. (2003) found datasets with as much as 92% missing data can still provide accurate phylogenetic reconstructions. When different genes have been used extensively in different sections of the tree (i.e., different genes for different taxonomic groups) and only a minimally overlapping set of species is available, a synthetic approach known as “supertree construction” must then be used (Keeling et al. 2005). The technological challenges and processing power required when dealing with datasets from large sequence databases is an area of ongoing research (e.g. Sanderson and Driskell, 2003; Beiko et al., 2005).

Phylogenetic inferences and species membership hypotheses generated from sequence database entries are only as good as the data on which they are based. Poor coverage of diversity in databases has already been discussed above, but what about reliability of the sequences already there? Most journals will require submission of sequences used in research articles to GenBank, although quality control of raw data is often solely dependent on the original scientists (Harris, 2003). Researchers are not obliged to correct existing database entries or use approved nomenclature (taxonomic and gene region), unless as a requirement of journal publication. Consequently, scientists including sequences from databases need to be aware that the quality is not always optimal, and should check unusual sequences in both a phylogenetic and functional context (Harris, 2003). Providing electropherograms to databases alongside the edited sequence read and not relying on computer base calls, opens up the actual sequence trace for scrutiny. Some apparent misidentifications may simply reflect disparities in taxonomy between specialists or knowledge of the systematics of the organism at the time (Bridge et al., 2004).

Case study: identities of the *Taygetis* butterflies of Costa Rica

An extensive ongoing inventory for all lepidopteran species in Area de Conservacion Guanacaste (ACG), Costa Rica was initiated in 2004 and incorporated DNA barcoding of selected specimens (Janzen et al. 2005, 2009; Janzen and Hallwachs, 2009).
When a high intra-specific sequence divergence was observed between two clusters of *Taygetis andromeda* (Cramer, 1776) (3% K2P vs. < 0.5%, the average conspecific K2P distance reported by Hajibabaei et al., 2006), a common satyrine recorded throughout much of South and Central America (DeVries, 1987), this prompted investigations into the identity of the two clusters and initiated a broader review of the taxonomic status of congeneric species known from Costa Rica. Of the roughly 27 species in the genus (Murray, 2001) DeVries (1987) lists 11 in Costa Rica, five of which are known from ACG (Janzen and Hallwachs, 2009).

After literature surveys, examination of images of types specimens, screening publicly available DNA sequences (GenBank) collected for other inventories, and study of diagnostic wing characters (Wilson, 2009), it was determined that of the 11 scientific names for *Taygetis* butterflies used by DeVries (1987), only three are still in valid use: *T. virgilia*, *T. kerea*, and *T. mermeria* (Figure 2). *T. andromeda* was determined to be a complex of two species: *T. laches* and *T. thamyra* (Janzen et al., 2009; Figure 2).

Incorporation of DNA barcodes into the species inventory proved effective at flagging cryptic diversity and assisting the resolution of issues in species-level taxonomy (see also Hebert et al., 2004; Vaglia et al., 2007; Floyd et al., 2009; Hausmann et al., 2009; Wilson et al., 2010). Barcodes were also crucial for simply checking identities of sequences used for phylogenetic studies and submitted to GenBank (Regier et al., 2009; Wilson, 2010). Figures 3 and 4 shows the Linnaean names attached to the *Taygetis* sequences on GenBank and the names currently applied by the sequence authors (e.g. Peña, 2011).

Given the rampant instability of the names encountered in this study, it is hard to imagine a situation where combinations will not change, or where previously overlooked synonymies will not become apparent. For example, phylogenetic analysis of DNA barcodes suggests Miller (2004) may have unwittingly created two paraphyletic genera when he followed the suggestion of Freitas (2003) and moved *Taygetis celia* to *Taygetomorpha* (Figure 3). Admittedly the study of *Taygetis* is incomplete, with a narrow taxonomic and geographical focus and many species still have an incomplete chain connecting the “name” to a barcode (Figure 2). However, this may not be cause for alarm. Instead, the barcodes offer new hope as independent taxonomic anchors, “taxonomy objects” *sensu* Vogler and Monaghan (2007) around which to investigate, connect and partition biodiversity.

**The future – DNA taxonomy?**

Advocates of DNA taxonomy have great ambitions for DNA sequence databases. Their vision is of a universal DNA-based taxonomy across all organismal groups with DNA sequences providing precise, digital descriptions and the “scaffold” for a classification system (Tautz et al., 2003). DNA taxonomy is founded on the reality that evolutionary entities i.e. taxonomic (phylogenetic) groups including species, are equally as distinguishable with DNA sequences as with morphological characters. If this future of taxonomy is to be realized, build-up of sequence databases is essential for the identification and classification of organisms (Savolainen et al., 2005). However, DNA taxonomy is
Figure 2. Diagnostic characters for ten *Taygetis* species from Costa Rica. Species of *Taygetis* found in Costa Rica presented in order of DeVries (1987) showing the connection between scientific names locked to type specimens and DNA barcodes through intermediary recently collected specimens. See Wilson (2009) for full details. *Taygetis mermeria*: (a) the type of *Taygetis excavate*; (b) ACG: 04-SRNP-11561; (c) BOLD: MHAAA177-05. *Taygetis virgilia*: (a) *Taygetis virgila rufomarginata* Lectotype 890329-18 (ZMHU); (b) ACG: 03-SRNP-17575; (c) BOLD: MHAAA737-05. *Taygetomorpha celia*: (a) the type of *Taygetis...
firmly rooted in the long-standing principles of traditional taxonomic work and reliant on names originating from previous classifications, otherwise new phylogenetic groupings would be incomprehensible to us (Franz, 2005). Vouchering specimens, including types and with the addition of DNA extractions, remains just as important a component as in the traditional approach.

Sequence information is easy to obtain, unambiguous, and makes species identification possible by non-specialists unfamiliar with the intricacies of morphology (Hebert et al., 2003a). Molecular operational taxonomic units (Blaxter, 2004), good species or not depending on the species concept applied, are nevertheless a good surrogate for identifying units of diversity in biodiversity studies. This enables users to obtain the information much faster than with the traditional morphological taxonomic process making surveys scalable across much larger taxonomic groupings and wider geographical regions (Smith et al., 2005) including correlation of “morphospecies” across diverse locations. It has taken two centuries to describe 1.7 million species using the traditional taxonomic approach. DNA-based species discovery has the potential to rapidly accelerate this process, an advantage which cannot be ignored in the light of the current biodiversity crisis affecting our planet. DNA sequences are already the de facto universal communication tool, providing data points for information about taxonomic

<table>
<thead>
<tr>
<th>Scientific Name</th>
<th>Specimen Details</th>
<th>Barcode Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taygetis kerea</td>
<td>Lectotype 091677A-19 (van Lennep collection; BMNH)</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis laches</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHACA63-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis thamyra</td>
<td>ACG: 04-SRNP-18609</td>
<td>BOLD: MHAA565-05</td>
</tr>
<tr>
<td>Taygetis kerea</td>
<td>ACG: 04-SRNP-13986</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis uzza</td>
<td>ACG: 06-SRNP-44842</td>
<td>BOLD: MHAC383-07</td>
</tr>
<tr>
<td>Taygetis sosis</td>
<td>from Weymer (1910)</td>
<td>BOLD: MHAC383-07</td>
</tr>
</tbody>
</table>

Figure 2. (Continued)
Figure 3. COI MP tree for *Taygetis* and allies. A single most parsimonious cladogram (Tree length 1230, CI 0.321, RI 0.712) was found by a heuristic search of a datamatrix containing 650 bp COI “DNA barcodes” from *Taygetis* species and allies available on GenBank or BOLD. Species of *Taygetis* listed in DeVries (1987) and present on the tree are highlighted in orange boxes. All species represented by more than one sequence are “monophyletic” on the cladogram. Where sequences are labelled with two names (marked with a red star) these refer to the names attached to the sequences on GenBank and the names currently applied by the sequence authors. In one case (blue star) it was impossible to confirm the identity of the specimen due to images of the specimen not being made available by the sequence author.

Figure 4. Wingless MP consensus tree for *Taygetis* and allies. The tree is a strict consensus tree of 29 most parsimonious cladograms (Tree length 386, CI 0.630, RI 0.701) found by a heuristic search of a datamatrix containing wingless gene fragments (400 bp) from *Taygetis* species and allies available on GenBank or newly generated for this study. Species of *Taygetis* listed in DeVries (1987) and present on the tree are highlighted in orange boxes. The two species represented by more than one sequence are “monophyletic” on the cladogram. Where sequences are labelled with two names (marked with a red star) these refer to the names attached to the sequences on GenBank and the names currently applied by the sequence authors. In one case (blue star) it was impossible to confirm the identity of the specimen due to images of the specimen not being made available by the sequence author.
specimens and species, in a way that complicated, mostly incomprehensible and unobtainable morphological descriptions can never be.

To some extent DNA taxonomy is already reality above the species level with systematists predominantly employing sequence data (and sequence databases) in phylogenetic reconstructions. The failure of researchers to translate molecular phylogenetic analyses into traditional classifications, and whether this is in fact necessary and useful, is discussed by Franz (2005). Sequence data collected for lots of different kinds of studies and purposes and submitted to sequence databases can be drafted into attempts to reconstruct the Tree of Life (Driskell et al., 2004).

In my view, taxonomy and DNA sequence databases are a good match for the following compelling reasons; DNA sequences are unambiguous digital data not influenced by subjective assessment and open to repeated analysis and testing of the species and phylogenetic hypotheses generated. Once submitted to online databases nucleotide sequences represent a freely available taxonomic resource that democratizes the way species are recognized and opens up biodiversity to all users.

Acknowledgements

Bob Hanner (University of Guelph) helped improve an earlier version of this manuscript. Dan Janzen, Winnie Hallwachs (University of Pennsylvania), Lee Miller, Jackie Miller (Florida Museum of Natural History), Mehrdad Hajibabaei and Paul Hebert (University of Guelph) are co-authors of the case study - We thank parataxonomists for finding and rearing the specimens from the ACG, the BOLD team (Megan Milton and Kara Layton) and CCDB laboratory staff (Suresh Naik) for technical support, Heather Braid (University of Guelph) for help while learning how to create a scratch-pad, and Niels Kristensen (Denmark), John Chainey (BMNH), Andre Freitas (Brazil) and Niklas Wahlberg (University of Turku) for help with taxonomic investigations.

References


