JOURNAL OF CRUSTACEAN BIOLOGY, 29(1): 96–110, 2009

"COI-LIKE" SEQUENCES ARE BECOMING PROBLEMATIC IN MOLECULAR SYSTEMATIC AND DNA BARCODING STUDIES

Jennifer E. Buhay

University of South Carolina, Belle W. Baruch Institute of Marine and Coastal Sciences, Columbia, South Carolina 29208, U.S.A. and IAP World Services, NOAA, 705 Convent Avenue, Pascagoula, Mississippi 39567, U.S.A. (jenbuhay@gmail.com)

ABSTRACT

The cytochrome c oxidase subunit I (COI) gene plays a pivotal role in a global effort to document biodiversity and continues to be a gene of choice in phylogenetic and phylogeographic studies. Due to increased attention on this gene as a species' barcode, quality control and sequence homology issues are re-emerging. Taylor and Knouft (2006) attempted to examine gonopod morphology in light of the subgeneric classification scheme within the freshwater crayfish genus *Orconectes* using COI sequences. However, their erroneous analyses were not only based on supposed mitochondrial sequences but also incorporated many questionable sequences due to the possible presence of numts and manual editing or sequencing errors. In fact, 22 of the 86 sequences were flagged as "COI-like" by GenBank due to the presence of stop codons and indels in what should be the open reading frame of a conservative protein-coding gene. A subsequent search of "COI-like" accessions in GenBank turned up a multitude of taxa across Crustacea from published and unpublished studies thereby warranting this illustrated discussion about quality control, pseudogenes, and sequence composition.

KEY WORDS: cytochrome c oxidase subunit I, molecular taxonomy, numt, protein-coding gene, pseudogene

DOI: 10.1651/08-3020.1

INTRODUCTION

The mitochondrial protein-coding gene, cytochrome c oxidase subunit I (COI), is a widely accepted marker for molecular identification to the species level across diverse taxa (examples of large scale projects: springtails, Hogg and Hebert, 2004; butterflies, Hebert et al., 2004a; birds, Hebert et al., 2004b; fishes, Ward et al., 2005; crustaceans, Costa et al., 2007). Approximately 700 nucleotides of COI molecular sequence can be used to query large COI datasets to help determine species' identity of unknown samples, a method known as "barcoding" (Hebert et al., 2003a, b). It is now possible to submit a sequence of unknown origin to the Consortium for the Barcoding of Life website (BOL: http://www.barcodinglife.org/views/idrequest.php) and within seconds, either the name of the species (if there is a reference sequence from that species accessioned into the database) or the name of the closest related taxa (if there is no reference sequence for species' comparison in the database) will appear on the query screen along with percent COI sequence similarity of the top 20 species' matches.

While the method of matching unknown molecular sequences to an online database is not new (for example, similar queries can be done with the Blast Search option in GenBank: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi), barcoding and similar methods using "molecular taxonomy" (Dayrat, 2005) such as "DNA Surveillance" (http://www.cebl.auckland.ac.nz:9000/) are highly dependent on 1) accurate identification of species in the reference database for comparison (http://www.barcoding.si.edu/DNABarcoding.htm) and 2) accurate molecular sequences. The backbone of the BOL relies heavily on the gathering of molecular data from preserved and curated museum specimens (vouchers), representing a collaboration between members of the BOL Consortium including among others, the National Museum of Natural History (Smithsonian Institution: www.si.edu), and the National Institutes of Health's online repository GenBank (http://www.ncbi.nlm.nih.gov/Genbank).

Ultimately, the responsibility of accurate identification of animal specimens rests with the researchers who determine species' identity using a host of morphological characters, and hopefully, these researchers create photo-documentation and/or accession museum vouchers to enable cross-checking of their species' diagnosis by others. While some physical characters are "better than others" to place a species' label on an organism, morphologically cryptic species remain cryptic species without examination of non-morphological features such as genetic sequence data. Morphological features are sometimes useless and misleading when trying to determine the species' identity of various larval stages, females for which keys are virtually non-existent in many animal groups, or specimens mutilated from intensive collection methods (such as trawls). Similarly, diagnostic morphological characters are sometimes missed, such as internal anatomy which can be difficult to dissect out or color patterns which are lost in the preservation process. These are all issues driving traditional taxonomists to move beyond the strict diagnoses of species, our units of biodiversity, using solely morphological information and into the realm of modern molecular approaches for more robust species diagnoses (Paquin and Hedin, 2004; Sites and Marshall, 2004).

The ease and low cost of gathering molecular data has made it rather commonplace for systematic biologists to sequence a standard set of genes for phylogenetic studies, particularly mitochondrial genes with "universal" primers,