Zootaxa 3609 (6): 593–600
www.mapress.com/zootaxa/

Article

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

http://dx.doi.org/10.11646/zootaxa.3609.6.5
http://zoobank.org/urn:lsid:zoobank.org:pub:E8AD9AD9-C6E4-4B73-BAA1-B25CE16F39E9

Publishing large DNA sequence data in reduced spaces and lasting formats, in paper or PDF

ALEXANDRE PIRES AGUIAR
Universidade Federal do Espírito Santo, Departamento de Ciências Biológicas, Av. Fernando Ferrari 514, Goiabeiras, Vitoria, ES, 29075–010, Brazil. E-mail: aguiar.2@osu.edu

Abstract

Scientific publications carry a practical moral duty: they must last. Along that line of thinking, some methods are proposed to allow economically and structurally viable publication of DNA sequence data of any size in printed matter and PDFs. The proposal is primarily aimed at contributing for preserving information for the future, while allowing authors to avoid information splitting and complement storage *ex situ*, that is, in server machines, outside the publication proper. The technique may also help to solve the impasse between the ICZN *Code* requirement that a new nomen be associated to diagnostic characters for the taxon *vs*. the phylogenetic definition of taxa, based on cladograms only: sequence data are characters, and can now be easily and comfortably included in taxonomic publications, with direct textual mention to their diagnostic sections. The compression level achieved allows the inclusion of all wanted DNA or RNA sequences in the same printed matter or PDF publications where the sequences are cited and discussed. Reduced font sizes, invisible fonts, and original 2D black & white and color barcodes are illustrated and briefly discussed. The level of data compression achieved can allow each full page of sequence data, or about 5000 characters, to be precisely coded into a color barcode as small as a square of 1.5 mm. A practical example is provided with *Taeniogonalos woodorum* Smith (Hymenoptera, Trigonalidae). Free software to generate publishable barcodes from txt or FASTA files is provided at www.systaxon.ufes.br/dna.

Key words: base pairs, COI, data compression, permanence, publication, RNA, taxonomy

Introduction

The publication of DNA or RNA sequences is problematic in its essence, due to some practical reasons. First, such sequences are usually large, counting hundreds or thousands of base pairs for each specimen or taxon treated in a paper. If printed with regular font size, this will usually lead to many extra printed pages, which is both cumbersome and often prohibitively expensive for editors and authors. Second, such sequences are primarily meant to be read or processed by machines, not humans, so it makes little sense printing them as conventional text. In fact, several initiatives already try to solve the need to store sequences, the GenBank and the Barcoding of Life Database (BOLD) currently representing some of the most widely known.

Printed text, however, remains as the most stable form of publishing information aimed to last. It still represents the nearest we can get of providing an accessible and permanent format for information, with nearly two thousand years of proved efficiency. Such permanence derives both from the fact that physical copies are produced, and from the fact that, currently, thousands of identical copies are distributed and stored throughout the world. Server machines, on the other hand, are often few for each publication, and are clearly and immesurably more fragile or unstable than paper, both physically and logistically. It seems therefore obvious that online documents are still far from achieving a similar degree of permanence to that of paper publications.

The permanence of electronic *vs*. paper publishing is however a heated debate, and it is not the aim of this work to review such extensive discussion—while self-sufficient, all techniques proposed herein can also be advantageously used *in addition* to the currently available options for storing molecular data, with gains for all. The reader can however find more in depth discussions directly or indirectly related to the permanence of electronic publications and internet links in Dubois (2003, 2010, 2011), Dimitrova & Bugeja (2007), Carlos & Voisin (2009),

Accepted by A. Dubois: 1 Nov. 2012; published: 4 Feb. 2013