Contributions to Zoology, 83 (1) 79-83 (2014)

ExcaliBAR: a simple and fast software utility to calculate intra- and interspecific distances from DNA barcodes

Mansour Aliabadian[1,6], Vincent Nijman[2], Ahmad Mahmoudi[1], Mehdi Naderi[3], Ronald Vonk[4], Miguel Vences[5]
[1] *Department of Biology, Faculty of Sciences, Ferdowsi University of Mashhad, Mashhad, Iran*
[2] *Oxford Brookes University, School of Social Sciences and Law, Department of Anthropology and Geography, OX3 0BP Oxford, United Kingdom*
[3] *Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran*
[4] *Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, The Netherlands*
[5] *Zoological Institute, Technical University Braunschweig, Mendelssohnstr. 4, 38106 Braunschweig, Germany*
[6] *E-mail: aliabadi@um.ac.ir*

Key words: bioinformatics, DNA Barcoding, pairwise genetic distance, sequence renaming

Abstract

In the context of DNA Barcoding, sequences of standard marker genes for thousands and potentially millions of individuals and species are becoming available, requiring ever more efficient bioinformatic environments and software algorithms for analysis. We here present ExcaliBAR (*Ex*traction, *Cal*culation, *Bar*coding), a user-friendly software utility to facilitate one important initial step in DNA barcoding analyses, namely the determination of the barcoding gap between pairwise genetic distances among and within species, based on original distance matrices computed by MEGA software. In addition, the software is able to rename sequences downloaded via the standard user interfaces of public databases such as GenBank, without the need of developing and applying specific scripts for this purpose.

Contents

Introduction 79
The program: DNA Barcode Distance Calculator (ExcaliBAR) 80
Automatic conversion and manual editing of sequence names 80
Format Converter 81
Availability and requirements 81
Environment 81
Instructions: loading, editing, and analysing data 81
Acknowledgements 83
References 83

Introduction

Many applications in bioinformatics deal with sequence analyses and sequence comparisons. The molecular revolution is providing DNA sequence data at an unprecedented and ever-increasing rate, posing challenges to molecular analyses. Several main types of such challenges can be distinguished (Sanderson and Driskell, 2003). On one hand, in phylogenomics, huge numbers of different orthologous gene sequences need to be extracted from genome data sets, aligned, the alignments quality-checked, and phylogenetic trees reconstructed (*e.g.* Delsuc *et al*., 2005). On the other hand, in the context of DNA Barcoding, sequences of standard marker genes for thousands and potentially millions of individuals and species are becoming available, requiring ever more efficient bioinformatic environments and software algorithms for analysis (*e.g.* Westram *et al*., 2011; Liu *et al*., 2012). DNA Barcoding with short and standardized DNA sequences of the mitochondrial gene cytochrome c oxidase subunit I (COXI) has been established as a fast, reliable and cheap method for species identification and discovery in animals (*e.g.* Hebert *et al*., 2003). Within the barcode of life project a tremendous amount of COXI sequences has been assembled. As of November 2013 more than 300,000 Barcode index numbers have been created for animals alone. The Barcode Index Number (BIN) System clusters sequences using well established algorithms to produce operational taxonomic units that closely correspond to species. In many cases, sets of data analysed in the context of DNA barcoding are collections of sequences from public databases such as EMBL Nucleotide Sequence Database (EMBL), The National Center for Biotechnology Information (NCBI), DNA Data Bank of Japan (DDBJ), or Barcode of Life Data Systems (BOLD). Although the main purpose of DNA barcoding is species identification, these sequences also

are regularly subjected to analysis with phylogeny software, to visualize sequence similarities or divergences or in some cases, to recover the relationships among closely related taxa. During our own studies (*e.g.* Vences *et al*., 2005; Aliabadian *et al*., 2009; Nijman and Aliabadian, 2010) we noticed that no simple bioinformatic tools were available to facilitate one important initial step in DNA barcoding analyses: the computation of pairwise genetic distances among and within species from sets of large numbers of sequences, as a basis for calculating the 'barcoding gap' between intraspecific and interspecific divergences in specific groups of organisms. It is the purpose of this paper to introduce a new software utility, named ExcaliBAR (Extraction, Calculation, Barcoding), that addresses parts of the processes in extraction (moment of sample determination), calculation and barcoding. It is adapted to make calculating the barcoding gap easy and straightforward to any end user. There are other software packages available that also can calculate DNA barcoding gaps beside their main function. For example; MEGA 4 or 5 (Tamura *et al*., 2007, 2011), the SPecies IDentity and Evolution package for R (SpideR; Brown *et al*., 2012), the Species Delimitation plugin for Geneious (Masters *et al*., 2011), the Automatic Barcode Gap Discovery (ABGD) website and Unix software (Puillandre *et al*., 2012), or the 'Barcode Gap Analysis' option on BOLD (http://boldsystems.org/), of which MEGA is the most practical one. However, these packages are not specifically devised to handle large data sets. The ExcaliBAR software does have this capability which is one of its main advantages. Further, the Species Delimitation plugin for Geneious calculates the inter- and intraspecific distances based on the nodes of reciprocal monophyly under the null model of random coalescence, therefore, there is no possibility to choose the substitution models for calculation of pairwise distances. But perhaps the most important advantage of the ExcaliBAR software is the relative ease at which it operates compared to other programs like SpideR for which the user needs to have basic knowledge of R to run the command. In addition ExcaliBAR uniquely provides an easy facility to transform long sequence titles from databases into short titles for subsequent phylogenetic and evolutionary analyses. ExcaliBAR is a user-friendly software utility that performs DNA Barcoding gap calculations after appropriately renaming sequences downloaded via the standard user interfaces of public databases such as NCBI, without the need of developing and applying specific scripts. Our utility transforms the titles of original Fasta file sequences from NCBI or BOLD to standard short titles, and subsequently calculates the interspecific and intraspecific pairwise differences, from a distance matrix previously calculated with MEGA 5 (Tamura *et al*., 2011).

The program: DNA Barcode Distance Calculator (ExcaliBAR)

The program has two parts, each with different functions. The first part edits the sequence names; in this part the program adjusts the title of different sources (from NCBI/EMBL/DDBJ or BOLD) into a uniform title. The second part is the Format Converter; in this part the distance matrices calculated by MEGA 4 or 5 (Tamura *et al*., 2007, 2011) are converted into two different files containing pairwise inter- and intra-specific distances. These distances might subsequently be used to calculate the barcoding gap, and to define a distance threshold as initial indicator for possible species boundaries and the identification of candidate species, which then can be the target of more focused taxonomic research (Vieites *et al*., 2009; Padial *et al*., 2010).

Automatic conversion and manual editing of sequence names

The name of sequences as they are routinely downloaded from different sources (mainly from NCBI and

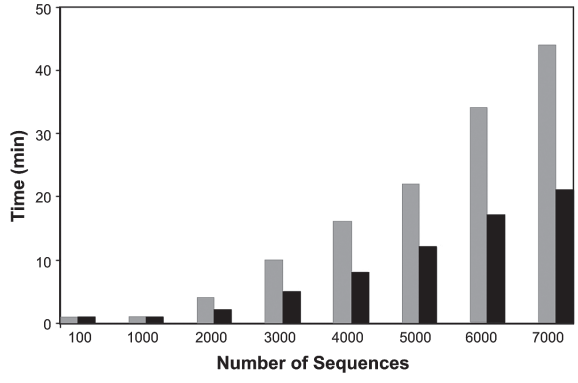

Fig. 1. Time requirements for loading the matrix file of distances from MEGA and assembling output of interspecific distances files of different sizes by Format Converter. Grey bars: minutes used for loading input files. Black bars: minutes used for assembling interspecific distance files. The different sizes are composed of aligned CO1 sequences with a length of 650 bp.

BOLD) are very long. Names of this length pose several problems. First, these names are used to identify the sequences on the tree that a phylogeny program eventually draws. By default many of these phylogeny programs like MEGA only use up to the first forty characters of the sequence description as the name. Second, some file formats (*e.g.*, Nexus) and some analysis programs using these do not permit characters other than letters and digits. Characters such as – ( ) * |, etc. therefore might raise problems and some editing of the sequence name is needed before an analysis can be performed. The ExcaliBAR software is capable to read the title of retrieved sequences from NCBI or BOLD SYSTEM in 'Fasta' format, and then edits the title of sequences into a readable format for further statistical analysis. On a double core Pentium P8700 PC with Windows 7 operating system, batch-renaming a file with 5000 sequences only takes about 60 seconds. Providing a unique name for each sequence, the program only keeps the accession or identification numbers of each sequence plus the genus, species, and subspecies names. These four names are separated by underscore characters (for example DQ683504_*Oenanthe_oenanthe_libanotica*). The program supports the Fasta format of the NCBI/EMBL/DDBJ and BOLD data sources although some names of available sequences are unconventional and cannot be edited by the program. In most cases the program highlights such problematic sequences, letting them appear with a different colour and allowing for their manual editing. The misidentified species, interim names, synonyms, and typos in sequences should be checked for before running the data with ExcaliBAR software. Problematic sequences will be flagged in inter- or intraspecific distances files as they will be strongly deviating.

Format Converter

The second part of ExcaliBAR is Format Converter. This part needs an input matrix file from MEGA 4 or 5 software (Tamura *et al*., 2007, 2011), in which the pairwise molecular distances have been calculated for sequences. The output of 'Compute pairwise' option in MEGA 4 or 5, saved in simple text format, can then be used as 'text' input file for the 'Format Converter' function. There are two options in the 'Format Converter' window that give the possibility to calculate and save intra- or interspecific pairwise distances. On a dual core Pentium P8700 PC with Windows 7 operating system, a Fasta file with 5000 *COX1* sequences (650 bp) takes about 22 minutes to load the input matrix, and 12 minutes to assemble a file with interspecific distances (Fig. 1). The result files can be saved either in Text or Microsoft Excel format. This option will be particularly useful for DNA barcoding applications that require computing inter- and intra-specific pairwise distances to flag the threshold above which sequences are likely to represent different species, and as such appear to be good candidates for further taxonomic research .

Availability and requirements

The open source software tool described herein, implementing the algorithms and models described in this paper, is freely available from www.um.ac.ir/~aliabadi/; www.mvences.de; and www.vincentnijman.org. The programme has also been deposited in the Dryad repository (doi:10.5061/dryad.r458n).
The ExcaliBAR software runs under Microsoft Windows, with Microsoft Net Framework version 2 or higher which is pre-installed in Microsoft Windows 7. Otherwise it can be downloaded and installed in the . NET Framework 4 from Microsoft website. The .NET Framework needs up to 500 MB of free disk space. The software runs only under Windows. It has been tested with Windows XP, Windows Vista and Windows 7, but presumably it should work also with Windows 2000 and Windows 8. Please report any problem with other Windows versions. To work with this program the ExcaliBAR.zip file only needs to be unzipped in any folder. No further installation is required.

Environment

ExcaliBAR software has a simple environment with three options: 'File, Tools and Help'. In addition, there are two shortcut icons under the titles menu bar, 'Start Edit' and 'Format Converter' that can be used to open input Fasta files and distances matrices (Fig. 2).

Instructions: loading, editing, and analysing data

Edit the sequence names

- To edit sequences names, select editor option under 'Tools', or use 'Start Edit' icon, to open the ExcaliBAR window. By browsing the input Fasta file from a given address, the original sequences of Fasta file will appear in the first column or 'Input sequences' section of the ExcaliBAR window. The output file, after applying the editing button, will appear in the second column or 'Edited title of sequences'. The

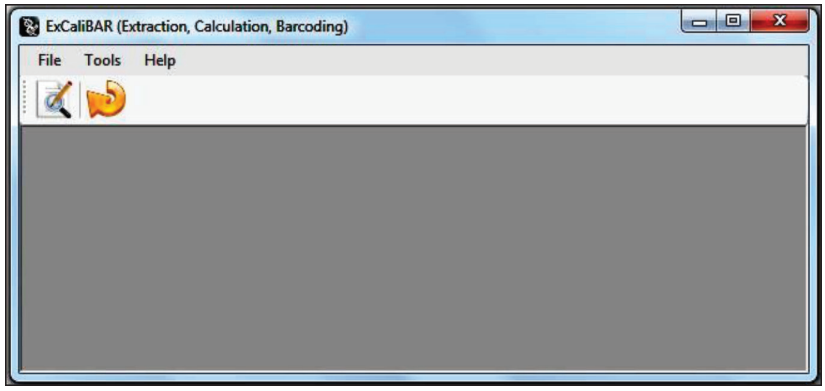

Fig. 2. The simple environment of ExcaliBAR software, including three options and two shortcut icons.

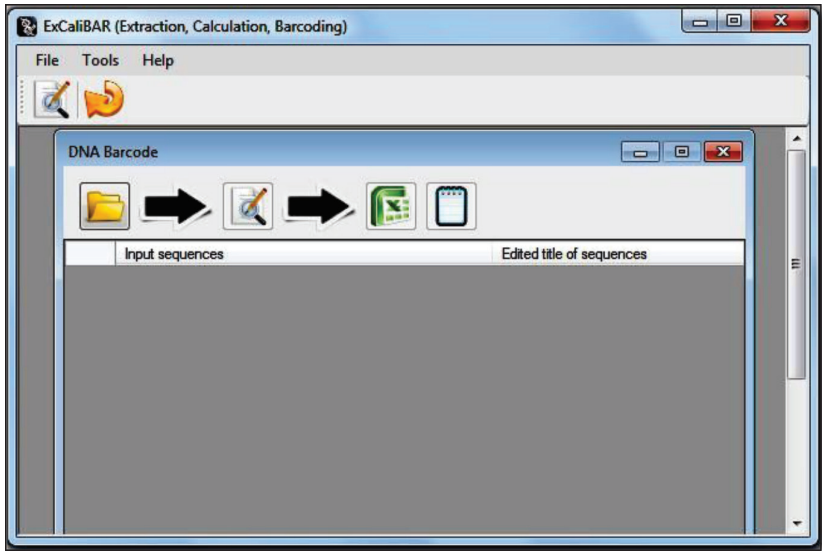

Fig. 3. The Edit window shows edited titles of sequences in the second column (edited title of sequences). The edited names including accession number_genus_species_subspecies name.

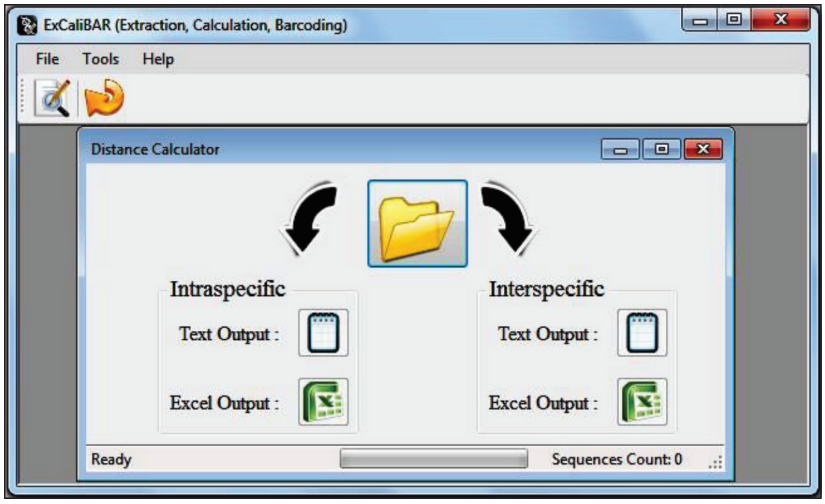

Fig. 4. Format Converter frame window for computing intra- and interspecific pairwise distances.

number of sequences read will appear in the menu bar below. The edited file can be saved either in text (Fasta) or excel format. At present some unconventional titles in GenBank or BOLD reside, and the software might not be able to recognize and edit the name of those sequences. In most of these cases, the title of the sequences will be marked with different colours in the 'Edited Genes' column. They can be edited manually (Fig. 3).

Format Converter

- Before starting with this part you need a computed pairwise distance matrix for your sequences. The matrix of general pairwise distances can be calculated with the 'Compute pairwise distances' option of MEGA 4 or 5. (Tamura *et al*., 2007). Keep in mind that assumed decimals for calculated distances are 4 digits and that the distance matrix should be saved in the lower left format. The text file of distances saved under MEGA (in simple text format) is then used as input file for the Format Converter.
- Open Format Converter under 'Tools', or click on 'Format Converter' icon in the tool bar menu, to open the ExcaliBAR window
- By clicking on the file icons, browse the computed pairwise distances text file as input file.
- There are two options in the right and left side of the file icons for calculating intra- and inter-specific pairwise molecular distances (Fig.4). The output of inter- and intraspecific distances could be saved as text or excel file. Excel 2007 is restricted to 1,048,576 rows, and if your comparison is above this figure, it is advised to save the output file, particularly for interspecific distances, in text format.
- You can see the progress of computing, and also the sequence counting on the lower part of the menu bar. Depending on the data size this computing takes different amounts of time.

Acknowledgements

This study was financially supported by a grant from the Zoological Museum Amsterdam, University of Amsterdam (to MA). We are indebted to Masoud Shirazian for writing the SPD 1.1 program and to an anonymous referee for comments on the manuscript.

References

Aliabadian M, Kaboli M, Nijman V, Vences M. 2009. Molecular identification of birds: performance of distance based barcoding in three genes to delimit closely related species. *PLoS ONE* 4: e4119.

Brown SD, Collins RA, Boyer S, Lefort MC, Malumbres-Olarte J, Vink CJ, Cruickshank RH. 2012. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources* 12: 562-565.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* 6: 361-375.

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

Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, Linder CR. 2012. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* 61: 90-106.

Masters BC, Fan V, Ross HA. 2011. Species delimitation – a geneious plugin for the exploration of species boundaries. *Molecular Ecology Resources* 11: 154-157.

Nijman V, Aliabadian M. 2010. Performance of distance-based DNA barcoding in the molecular identification of Primates. *Comptes rendus Biologies* 333: 11-16.

Padial JM, Miralles A, de la Riva I, Vences M. 2010. The integrative future of taxonomy. *Frontiers in Zoology* 7: 16.

Puillandre N, Lambert A, Brouillet S, Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology* 21: 1864-1877.

Sanderson MJ, Driskell AC. 2003. The challenge of constructing large phylogenetic trees. *Trends in Plant Science* 8: 374-379.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* 28: 2731-2739.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24: 1596-1599.

Vences M, Thomas M, Bonett RM, Vieites DR. 2005. Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society London*, *Ser. B*, 360: 1859-1868.

Vieites DR, Wollenberg KC, Andreone F, Köhler J, Glaw F, Vences M. 2009. Vast underestimation of Madagascar's biodiversity evidenced by an integrative amphibian inventory. *Proceedings of the National Academy of Sciences of the USA* 106: 8267-8272.

Westram R, Bader K, Prüße E, Kumar Y, Meier H, Glöckner FO, Ludwig W. 2011. ARB: a software environment for sequence data. Pp. 399-406 in: FJ de Bruijn, ed., Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches, Wiley-Blackwell.

Received: 17 April 2013
Revised and accepted: 9 December 2013
Published online: 18 February 2014
Editor: J.W. Arntzen