OPEN ACCESS Freely available online

PLoS one

VoSeq: A Voucher and DNA Sequence Web Application

Carlos Peña[1]*, Tobias Malm[2]

1 Department of Biology, University of Turku, Turku, Finland, **2** Department of Biology, University of Eastern Finland, Joensuu, Finland

Abstract

There is an ever growing number of molecular phylogenetic studies published, due to, in part, the advent of new techniques that allow cheap and quick DNA sequencing. Hence, the demand for relational databases with which to manage and annotate the amassing DNA sequences, genes, voucher specimens and associated biological data is increasing. In addition, a user-friendly interface is necessary for easy integration and management of the data stored in the database back-end. Available databases allow management of a wide variety of biological data. However, most database systems are not specifically constructed with the aim of being an organizational tool for researchers working in phylogenetic inference. We here report a new software facilitating easy management of voucher and sequence data, consisting of a relational database as back-end for a graphic user interface accessed via a web browser. The application, VoSeq, includes tools for creating molecular datasets of DNA or amino acid sequences ready to be used in commonly used phylogenetic software such as RAxML, TNT, MrBayes and PAUP, as well as for creating tables ready for publishing. It also has inbuilt BLAST capabilities against all DNA sequences stored in VoSeq as well as sequences in NCBI GenBank. By using mash-ups and calls to web services, VoSeq allows easy integration with public services such as Yahoo! Maps, Flickr, Encyclopedia of Life (EOL) and GBIF (by generating data-dumps that can be processed with GBIF's Integrated Publishing Toolkit).

Citation: Peña C, Malm T (2012) VoSeq: A Voucher and DNA Sequence Web Application. PLoS ONE 7(6): e39071. doi:10.1371/journal.pone.0039071

Editor: Jonathan H. Badger, J. Craig Venter Institute, United States of America

Received April 2, 2012; **Accepted** May 17, 2012; **Published** June 12, 2012

Funding: This work was supported by a grant from the Kone Foundation, Finland to Niklas Wahlberg and Tommi Nyman. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: carlos.pena@utu.fi

Introduction

The advent of molecular methods, such as DNA sequencing, has facilitated a rapid development of hypotheses for phylogenetic relationships among biological groups. The amount of DNA sequences that can be used in phylogenetic inference is growing at an increasingly fast pace. There are, as of now, more than 135 million sequences in GenBank [1] and 1.5 million DNA barcodes in BOLD [2]. Thanks to the development of new techniques and the possibility to outsource the processes involved, DNA sequencing is becoming cheaper and quicker. Thus, research groups might need the use of an efficient system in order to keep track of, manage and annotate their DNA sequences. In addition to adequate storage of DNA sequences, it would be advantageous if this system facilitates further processing of the data such as easy creation of datasets for phylogenetic analysis, quick creation of tables ready for publication and submission of sequences to public repositories (e.g. GenBank).

Relational databases permit organized storage of information that can be searched and retrieved quickly, depending on (and flexible to) the needs of users. However, relational databases such as MySQL need to be coupled with a software interface designed to act as a bridge between users and relational database for uploading, managing and retrieving e.g. DNA sequence and voucher specimen data.

Although there are several database systems available that handle biological and molecular data, they are conceived to be multipurpose systems [3]. Current biological databases handle a variety of information such as scientific names, current taxonomic classification, voucher images, geographic distribution maps, bibliographic references and so on. The Scratchpads project [4] is a social network portal for managing, sharing and publishing taxonomic information and voucher data online, but includes very limited capabilities for handling DNA sequence data (e.g. The Holometabola Insects Phylogenetic Database is based on Scratchpads [5]). The Lifedesks platform [6] hold taxonomic information and specimen data connected to voucher specimens in EOL, but there are no capabilities for storing and processing DNA sequences. The Mantis biological database manager allows storage of taxonomic and specimen data for museum collections [7]. These systems allow being configured so that they can used as private working platforms restricted to members of a research group.

The Barcode of Life Database (BOLD [8]) is both a public repository of DNA barcodes (and voucher data) and an online workbench module to collect and analyze DNA barcode sequences (a segment of the COI gene), but offers limited functionality for other gene markers as well as for dataset creation. Another commonly used database to store DNA sequences is GenBank, which is an excellent repository of data. However, GenBank is mainly used as a repository of sequences after they have been managed and analyzed during the research process. After the analysis of sequences is finished, the sequences are submitted to GenBank during the preparation of manuscripts. In this way, GenBank becomes a repository of end-products (sequences) of the research process.

Thus, there is a need for a database system that can be used to store and manage DNA sequences during the research process. This system should facilitate the aggregation of sequences into dataset files ready-to-run in common phylogenetic software. It should be able to handle the amount of data normally used during

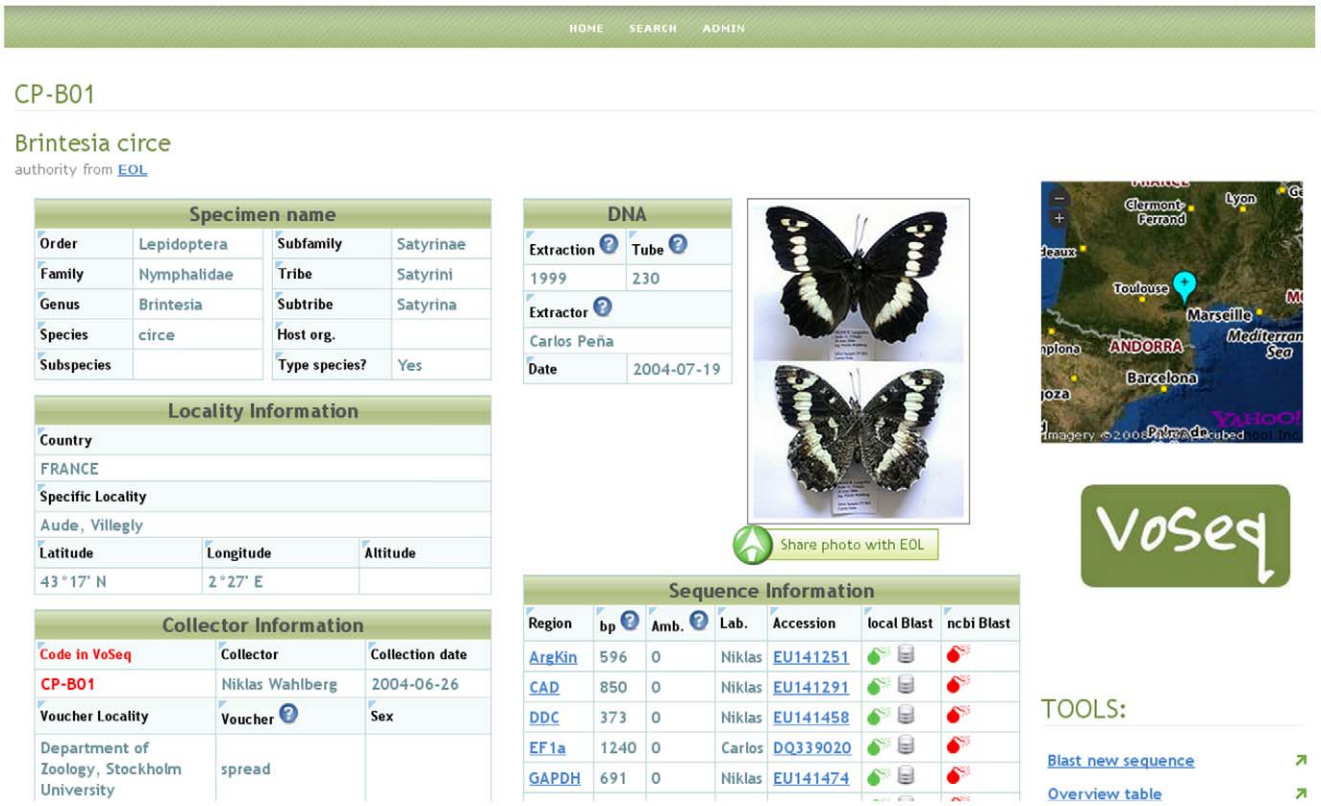

Figure 1. Screenshot of voucher page in VoSeq. It shows specimen and collection data, links to DNA sequences in GenBank, local and remote BLAST tools and mash-ups with Flickr and Yahoo Maps! for voucher picture and geographic location.
doi:10.1371/journal.pone.0039071.g001

phylogenetic research: such as biological data from hundreds of voucher specimens and thousands of DNA sequences for a number of gene markers.

Here we present VoSeq, a user friendly database system designed for day to day use by researchers in phylogenetic inference to store and organize DNA sequences. VoSeq also stores complementary data such as voucher photos, collection data, taxonomic classification and collection locality maps. VoSeq can be downloaded and installed by on-screen instructions on a private computer, or installed on a shared server so that it can be used as a web application restricted to a research group or a network of collaborators. The main functionality lies in retrieving ready-to-run datasets for phylogenetic analyses for various software, complete with partition tables and analysis specifications, as well as MS Excel tables for work overview or publications. The platform of this database is taxon independent and can be used for all organisms. The source code is freely available under the GNU General Public License v2. Additionally, VoSeq has been designed to be cooperative with other biological web-based databases. For this, we included features of the so-called Web 2.0 [9] such as the Ajax protocol for user-friendliness, and exchange of information over the Internet based on SOAP and REST calls using the XML and JSON formats.

Materials and Methods

VoSeq is written in the PHP scripting language to dynamically generate HTML and JavaScript code. All the information is stored in a MySQL relational database within several table trees. The database is designed to be used as a standalone application and accessed by using a standard web browser (e.g. Microsoft Internet Explorer, Mozilla Firefox or Google Chrome) provided that it is run on top of web server software (e.g. Apache [10]). If VoSeq is installed on a public server, it becomes a web server application, facilitating simultaneous collaboration between users from different geographic locations.

Results and Discussion

Design and implementation

After download, VoSeq is installed by following the on-screen instructions to complete the configuration process. VoSeq is platform independent and has been tested in Windows, Mac and Linux systems.

VoSeq consists of two interfaces, an Administrative interface for uploading and updating data, creation of user accounts, taxon sets and gene descriptions, and a User interface for data query and retrieval.

In the User interface, all available data for each voucher sample is summarized in its voucher page. When available, this page (Figure 1) shows taxonomic data, collection information along with an interactive map, a voucher picture and the list of DNA sequences including gene region, number of base pairs, accession numbers and a link to a "sequence page" where users can access the actual sequence and other information such as primers used in PCR amplification. Vouchers can be searched through an advanced search tool. Queries are made by searching either single or combinations of fields. Most of the search fields, as well as other entry fields within the database, are fitted with auto-complete drop-boxes for user-friendliness (based on Ajax).

The Administrative interface allows users to create new entries, upload and update all data and post voucher pictures that are hosted in the web service Flickr [11]. When users upload a picture in VoSeq, an algorithm will post the picture to Flickr and register the web addresses in MySQL. This facilitates showing the picture in the corresponding voucher page. This Flickr plug-in is enabled by following the instructions that will appear in VoSeq the first time a user tries to upload a picture. Once the user obtains an account in Flickr, VoSeq will instruct how to get a private key and how to register it in VoSeq's configuration file (this process needs to be done only once). New voucher and sequence records can be created in VoSeq by using the Administrative interface for single entries. Voucher codes in VoSeq are unique and cannot be overwritten, but can be changed. Moreover, VoSeq also includes a tool for uploading batches of voucher data and sequences. Data from a MS Excel sheet or other tab-delimited table can be copied and pasted into VoSeq and all the data will be processed and stored ready to be queried and retrieved. Records that are being used in phylogenetic projects can be included into "taxon-sets", either manually by marking the voucher in an overview table or added as a list, in the Administrative interface. Taxon-sets allow easy querying of voucher and sequence information or creation of datasets to be analyzed in phylogenetic software.

VoSeq is password protected and at installation open only to the installer, but additional user accounts are easily created, with or without administrative rights.

Data retrieval

One key feature of VoSeq is the possibility to harvest batches of DNA sequences for phylogenetic analysis. Users can create datasets consisting of selected DNA (or amino acid) sequences including flexible choices of outgroup and ingroup taxa, gene markers, codon positions and taxon-sets as well as different partitioning schemes. The chosen data are retrieved in ready-to-run datasets in NEXUS [12], TNT [13] and PHYLIP [14] file formats to be used, as-is or modified, as input to phylogenetic software such as MrBayes [15], TNT, RAxML [16] and others.

VoSeq also includes BLAST capabilities [17]. By following on-screen instructions, users can install NCBI BLAST software used for finding matching homologous sequences, for a new sequence or an already stored sequence, among those hosted in VoSeq. It is also possible to BLAST users' sequences against those hosted in GenBank.

VoSeq also includes a tool to create MS Excel tables with information on voucher specimens and available sequences, as well as accession numbers for sequences used for preparing a manuscript for publication. These tables are practically ready to be included in submission to journals. FASTA files appropriate for submission of chosen taxa and sequences to GenBank are also readily made with a few mouse-clicks. These fields include information such as organism name and lineage, gene codes and specimen voucher codes and can be imported into software such as Sequin [18].

Relation to other databases

In addition to BLAST capabilities against sequences in GenBank, VoSeq facilitates integration with other biological databases and public services available on the web. VoSeq is able to create a mash-up using Yahoo! Maps to plot the collection locality of vouchers. If the database has geographic coordinates for the particular voucher, its page will show the map with a tag pinpointing the specific locality for the voucher. This map can be zoomed in and out and dragged around by using the mouse. This is accomplished by implementing the Yahoo! Maps web service [19].

When users upload a voucher photo to VoSeq, it is automatically hosted in the user's Flickr account. From the voucher pages in VoSeq, by clicking the "Share with EOL" button, VoSeq will automatically submit voucher photos to the photo pool of the Encyclopedia of Life (EOL) [20] that is also hosted in Flickr.

VoSeq makes automated calls to EOL's web services [21] in order to pull information on authors and year of description for species. VoSeq sends genus and species names and waits for a response. If EOL response is positive, the full species name will be included in voucher pages.

VoSeq also facilitates sharing data with GBIF [22] by creating single-click data dumps that can be processed with their Integrated Publishing Toolkit (IPT), which is the method preferred by GBIF.

A very detailed error reporting system is used for minimizing downtime and making VoSeq as user-friendly as possible.

Availability and future directions

The software is open source with a GPL v2 license: https://github.com/carlosp420/VoSeq. A test installation with sample data can be found at http://www.nymphalidae.net/VoSeq. The full documentation and how-to help can be accessed here: http://nymphalidae.utu.fi/cpena/VoSeq_docu.html.

VoSeq is a useful web-based database application aimed for molecular systematics. VoSeq's simple and intuitive interfaces allow users to organize molecular sequences and related biological data. Its tools for retrieving batches of sequences in FASTA format allows them to be easily exported to specialized software, while the tool for creating PHYLIP, NEXUS and TNT ready-to-use datasets prove very time-saving.

VoSeq is entirely open source (available on Github at [23]) and savvy users can tweak the code to produce variants for subsets of data, as well as adding specific extra functions. Github is a convenient platform for sharing VoSeq's source code because it can be easily branched into new projects. For example, the source code could be made to call external programs for automatic analysis runs in wanted applications, or users can add extra fields to the database suitable to their organism group and research project, such as additional environmental variables or behavioral information. We are also planning to incorporate a fast Neighbor-Joining algorithm to the dataset section for a quick and easy look at hypothesized relationships among chosen taxa/taxon set. We do like to keep the software focused on the main functions, e.g. voucher and sequence data storage and retrieval, but users are free and welcome to suggest additions to or modification of the application.

As the Internet is currently the most important medium for delivery of data, biodiversity informatics needs tools that automate the exchange of data over the Internet in order to integrate biological information available from a disparate array of sources [24]. One way of solving this problem is by having loosely interconnected databases over the Internet so that all relevant data regarding species of interest can be aggregated on-the-fly and be presented to users in only one website (e.g. [25]). Thus, all digitized biological information will be readily available to people. All that is needed is (1) tweaking existent and new databases to provide data in format that computers understand (i.e. XML, JSON and variants) and (2) use of unique identifiers [26].

Acknowledgments

We extend our thanks to Niklas Wahlberg, Tommi Nyman, Marko Mutanen (University of Oulu, Finland), Marianne Espeland (Harvard University, USA) and Sanna Leppänen (University of Eastern Finland, Finland) for thoroughly testing as well as requesting features to be added to VoSeq. We acknowledge anonymous reviewers for comments on the manuscript.

Author Contributions

Conceived and designed the experiments: CP TM. Performed the experiments: CP TM. Analyzed the data: CP TM. Contributed reagents/materials/analysis tools: CP TM. Wrote the paper: CP TM. Developed the structure of VoSeq and basic functions: CP. Developed additional functions and tools: TM.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Research 39: D32–D37.
2. Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7: 355–364.
3. Smith VS, Rycroft SD, Brake I, Scott B, Baker E, et al. (2011) Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. ZooKeys 150: 53–70.
4. Scratchpads, biodiversity online. Available: http://scratchpads.eu/. Accessed 11 May 2012.
5. The Holometabola Insects Phylogenetic Database. Available http://hip-db.myspecies.info/hipdb/records. Accessed 11 May 2012.
6. Lifedesks. Available: www.lifedesks.org/. Accessed 11 May 2012.
7. Mantis. Available: http://140.247.119.138/mantis/. Accessed 11 May 2012.
8. Barcode of Life Database. Available: http://www.boldsystems.org. Accessed 11 May 2012.
9. Web 2.0. Available: http://en.wikipedia.org/wiki/Web_2.0. Accessed 29 March 2012.
10. Apache HTTP Server Project. Available: http://httpd.apache.org/. Accessed 11 May 2012.
11. Flickr. Available: http://www.flickr.com/. Accessed 29 March 2012.
12. Maddison DR, Swofford DL, Maddison WP (1997) Nexus: An extensible file format for systematic information. Systematic Biology 46: 590–621. doi:10.1093/sysbio/46.4.590.
13. Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24: 774–786. doi:10.1111/j.1096-0031.2008.00217.x.
14. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
15. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
16. Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Systematic Biology 57: 758–771.
17. BLAST. Basic Local Alignment Search Tool. Available: http://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed 29 March 2012.
18. Sequin. A DNA Sequence Submission and Update Tool. Available: http://www.ncbi.nlm.nih.gov/Sequin/. Accessed 20 May 2012.
19. Yahoo! Maps Web Services. Available: http://developer.yahoo.com/maps/. Accessed 29 March 2012.
20. EOL. Encyclopedia of Life. Available: http://www.eol.org. Accessed 29 March 2012.
21. EOL API. Encyclopedia of life. Available: http://eol.org/api/. Accessed 29 March 2012.
22. GBIF. Global Biodiversity Information Facility. Available: http://www.gbif.org/. Accessed 29 March 2012.
23. VoSeq source code. Available: https://github.com/carlosp420/VoSeq. Accessed 11 May 2012.
24. Parr CS, Guralnick R, Cellinese N, Page RDM (2011) Evolutionary informatics: unifying knowledge about the diversity of life. Trends in Ecology & Evolution 27: 94–103.
25. Page RDM (2011) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. BMC Bioinformatics 12: 187.
26. Page RDM (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9: 345–354.