BLOG 2.0: a software system for character‐based species classification with DNA Barcode sequences. What it does, how to use it

Emanuel Weitschek, Robin Van Velzen, Giovanni Felici, Paola Bertolazzi

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 1043–1046 doi: 10.1111/1755-0998.12073

BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it

EMANUEL WEITSCHEK,*† ROBIN VAN VELZEN,‡§ GIOVANNI FELICI* and PAOLA BERTOLAZZI*
*Institute of Systems Analysis and Computer Science A. Ruberti, National Research Council, Viale Manzoni 30, 00185 Rome, Italy, †Department of Informatics and Automation, Università degli Studi Roma Tre, Via della Vasca Navale 79, 00146 Rome, Italy, ‡Biosystematics Group, Wageningen University, Wageningen, the Netherlands, §Naturalis Biodiversity Center (section NHN), Wageningen University, Wageningen, the Netherlands

ABSTRACT

BLOG (Barcoding with LOGic) is a diagnostic and character-based DNA Barcode analysis method. Its aim is to classify specimens to species based on DNA Barcode sequences and on a supervised machine learning approach, using classification rules that compactly characterize species in terms of DNA Barcode locations of key diagnostic nucleotides. The BLOG 2.0 software, its fundamental modules, online/offline user interfaces and recent improvements are described. These improvements affect both methodology and software design, and lead to the availability of different releases on the website http://dmb.iasi.cnr.it/blog-downloads.php. Previous and new experimental tests show that BLOG 2.0 outperforms previous versions as well as other DNA Barcode analysis methods.

Keywords: classification, data analysis, DNA Barcoding, species identification

Received 6 June 2012; revision received 19 December 2012; accepted 22 December 2012

BLOG 2.0

The specimen classification technique named DNA Barcoding was proposed by Hebert *et al.* (Hebert *et al.* 2003). A short DNA sequence from a small portion of the mitochondrial DNA, the gene cytochrome c oxidase subunit I (*COI*), was chosen as Barcode for animals and, more recently, a combination of two different gene regions (*rbcL* and *matK*) was defined as Barcode for plants (CBOL Plant Working Group 2009); the internal transcribed spacer (ITS) gene region was proposed as a universal Barcode marker for Fungi (Schoch *et al.* 2012). These small portions of the DNA present high variability, also between closely related species, and are considered to contain sufficient information to classify a specimen to species. DNA Barcoding may, in certain contexts, be applied also to the more general problem of taxa classification; however, the types of Barcode adopted in this work and in the related literature have always been used for specimen classification at the species level of the phylogenetic tree.

Several data analysis methods have been developed and adopted to automatically classify a DNA Barcode sequence to a predefined species, such as tree-based methods, similarity-based methods and diagnostic methods. For a complete survey refer to.

(van Velzen *et al.* 2012). Most of these methods are available on online services, like http://bol.uvm.edu (Sarkar & Trizna 2011) and http://www.boldsystems.org (Ratnasingham & Hebert 2007).

The classification problem may be formulated in the following way: given a reference library composed of DNA Barcode specimen sequences of known species and an unknown DNA Barcode sequence, recognize the latter into a species that is present in the library.

In this application note we present version 2.0 of the character-based diagnostic DNA Barcode analysis method BLOG, which is an evolution of the logic data mining method already described in (Bertolazzi *et al.* 2009). BLOG identifies for each species in the reference library the distinctive nucleotide positions of the DNA Barcode sequences, and assigns to each species logic classification formulas – small rules in the form of "*if-then*" – that are able to characterize a species in a compact way. An example of a BLOG rule is

"*if pos40 = T and pos265 = T then the specimen is classified as Ompok bimaculatus*".

A distinctive advantage of BLOG compared with other available methods is that such logic formulas offer additional species-level information that can be used

Correspondence: Emanuel Weitschek, Fax: +39067716450;
E-mail: emanuel.weitschek@iasi.cnr.it

outside the scope of DNA barcoding, for example, in species description, in molecular detection (van Velzen *et al.* 2012) or in phylogenetic analysis.

BLOG is based on two main computational steps:

1 Feature selection: BLOG selects a small set of positions of the DNA Barcode sequences that are suited to distinguish among the species in the reference library.
2 Formula extraction: BLOG computes the logic formulas that classify each species present in the reference library.

BLOG uses a supervised machine learning approach: the user has to provide as input a training set containing specimens with *a priori* known species membership. Based on this training set, the software selects suitable nucleotide positions (feature selection) and computes the logic formulas for species classification (formula extraction). Subsequently, the logic formulas can be applied to a test set which contains specimens that require classification. The test set can contain query specimens with unknown species membership or, alternatively, specimens that also have *a priori* known species membership, allowing verification of the specimen classifications. BLOG is designed to identify the locations of key diagnostic nucleotides for each species in a fully defined training set: to obtain reliable results, the testing set has to contain only specimens from the same species that are present in the training set. Also, a complete reference library of polymorphisms for each species is required in the training set to avoid false negatives.

The main evolutions of BLOG 2.0 reside in the availability of enhanced user interfaces, in a new classification algorithm, in the re-engineering of the software, in the format of its output, and in an optimized selection criteria of the candidate distinctive nucleotide positions.

Input and output

Input files are DNA Barcode sequence in standard FASTA format (Pearson 1990). The sequences have to be of the same region or pre-aligned to the same region before being processed by BLOG (e.g. sub-segments of COI or rbcL).

Output of BLOG are logic formulas for species classification, classification rates and confusion matrices. The logic formulas are small 'if-then rules' which assign a specimen to the species. Classification rates are given as number and percentage of correct, incorrect and not classified specimens. Confusion matrices give detail information on classification accuracy and cross-classification. The i-j cell of the matrix represents the number of specimens from species i predicted to be of species j. Correctly classified elements are on the main diagonal of the confusion matrix.

Feature selection

The first computational step of BLOG is the extraction of species-specific positions of the DNA Barcode sequences from the training set. The feature selection approach of BLOG is based on the mathematical optimization formulation described in (Bertolazzi *et al.* 2010). This approach has proven efficient and effective in many applications, such as classification of biological sequences (Bertolazzi *et al.* 2009, 2010; Weitschek *et al.* 2011, 2012a,b; van Velzen *et al.* 2012), and the analysis of numerical data such as gene expression profiles (Arisi *et al.* 2011; Weitschek *et al.* 2012a,b). As shown in the cited references, the mathematical formulation of the feature selection problem is NP-hard and cannot be solved at optimality for large instances. BLOG adopts an effective heuristic algorithm based on randomized search that is able to produce solution of high quality in limited time (a feasible solution is produced in linear time in problem size). The solution time is driven by the number of iterations – a user defined parameter – and experimentally it was verified that for Barcode instances such parameter needs to grow linearly with problem size.

Previous versions of BLOG (Bertolazzi *et al.* 2009) applied the feature selection step simultaneously on all species in the reference database. However, features that allow separation of one species are not necessarily useful for separating another. BLOG 2.0 therefore can apply the feature selection step separately to each species in the reference library. In each feature selection step, the considered species is assigned class A and all the other species class B. Consequently, m different instances of the feature selection problem have to be solved for each analysis run, where m is the number of species in the training set. A large computation time would be needed with exact algorithms which further justifies the use of the GRASP heuristic.

Formula extraction

The aim of the formula extraction step is to produce a logic formula (or rule) separating each species. BLOG adopts the Lsquare method (Felici & Truemper 2002), where the extraction of logic formulas is obtained by the solution of a sequence of well-known and hard logic optimization problem in the form of Minimum Cost Satisfiability Problems (MinSat). An extensive explanation of Lsquare and on the MinSat formulation is available in (Felici & Truemper 2002, 2006). Each literal of a formula represents an assignment of a nucleotide (i.e. A,T,G or C) to a particular position in the DNA Barcode sequence.

Previous versions of BLOG commonly produced formulas with both positive and negative literals (e.g. *pos40 = NOT T*) to minimize formulas size. However,

negative literals recognize three different nucleotides making them potentially less precise than positive literals (e.g. pos40 = G OR pos40 = C would be a more precise formula than pos40 = NOT T). Therefore, BLOG 2.0 allows increasing the cost of the negative literals in the MinSat problem formulations to prevalently output positive literals.

Classification

Before evaluating the test set, BLOG 2.0 performs an evaluation of the training set with the aim to assign relative weights to the logic formulas, according to the algorithm described in (Weitschek *et al.* 2011): the Laplace Score (Tan *et al.* 2005), the false positive and true positive rates are computed for every logic formula over the reference library, these scores are then considered in the test set for performing the classification assignments.

A typical complete experimental run (consisting in 1000 specimens belonging to 50 species) with BLOG 2.0 requires less than five minutes on a standard desktop machine (Intel Core i5, 4GB RAM).

Releases

Three releases of BLOG 2.0 are available, Graphical user interface, Command-line interface and a Web release; they are described in detail below.

Graphical user interface

An offline graphical user interface release is available for download on http://dmb.iasi.cnr.it/blog-downloads.php. We suggest this release of BLOG 2.0 for most users, who wish to fine tune the analysis and run the software on their own computers (Linux and Windows) as it has the most user-friendly interface. Users can graphically view the DNA Barcode sequences, load training and test files, execute BLOG 2.0 and view the classification results and the logic formulas for each species present in the data set. The offline graphic user interface has been implemented with the Java Swing framework. A complete user manual for this version is provided in the BLOG-2-GUI-manual.pdf supplementary material file.

Command-line interface

For performing intensive experimentations, we suggest to use the offline command-line version, which is available for download at http://dmb.iasi.cnr.it/blog-downloads.php. With this version, the user can organize experiments in batches and read the output in different files for each run. Executables of the BLOG software are available for Linux and Windows, and the C source code is released for compilation on other operating systems. A complete user manual for this version is provided in the supplementary material file.

BLOG2-COMMAND-LINE-README.txt.

Web release

A simple web user interface of BLOG is available at http://dmb.iasi.cnr.it/blog.php. Data (training and test sets) can be uploaded through an input form and results the (classification rates, logic formulas and confusion matrices) are returned in CSV (Comma Separated Values) text files, which are easily readable by a common spreadsheet software. In addition, a compressed archive containing all results is sent to the user via email. We direct the users to http://dmb.iasi.cnr.it/blog.php for additional information and usage instruction for this release. The BLOG web service has been released on a Linux server (Ubuntu Server distribution), using a LAMP platform (Linux Kernel 2.6.32, Apache 2.2.14, PHP 5.2) with a Java job queuing system that relies on a MySQL database (v. 5.1.41).

Discussion and conclusions

The BLOG 2.0 system has already been experimentally tested on various data sets (COI, ITS) and accurately compared with other competing methods in (Weitschek *et al.* 2011) and in (van Velzen *et al.* 2012).

Weitschek *et al.* (2011) found BLOG 2.0 outperformed BLOG 1.0 based on three empirical DNA Barcode data sets (bats, fishes and birds, available on http://dmb.iasi.cnr.it/blog-downloads.php).

In van Velzen *et al.* (2012), a comparison of the relative performance of DNA Barcode data analysis methods in identifying recently diverged species was performed. The authors compared tree-based methods, similarity-based methods and diagnostic methods using simulated, as well as empirical DNA Barcode data sets (all available on http://dmb.iasi.cnr.it/blog-downloads.php). The diagnostic method BLOG had highest correct query identification rate based on simulated as well as empirical data, indicating that it is a consistently better method overall.

To consolidate the performance of BLOG, the software was tested on two new data sets, the first composed of internal transcribed spacer (ITS) gene region Barcode fungi sequences and the second containing ribulose-bisphosphate carboxylase gene (rbcl) region green algae Barcode sequences (both available on http://dmb.iasi.cnr.it/blog-downloads.php). In particular, 50 fungi sequences belonging to eight different species in the Dikarya subkingdom and 26 green algae

sequences of five different species in the Haematococcaceae family were extracted from BOLD (Ratnasingham & Hebert 2007). The results were in line with the classification rates obtained with previous experiments on COI and ITS sequences: for fungi 92% correct classification rates (sensitivity 0.923, specificity 1), for algae 100% correct classification rates (sensitivity 1, specificity 1) and compact classification formulas composed of one or two nucleotides locations.

Beyond the promising classification results, the distinctive advantage of BLOG is the output of the model, which gives a compact and precise description of species in the reference library. BLOG offers additional species-level information –the logic classification formulas – that may also be used outside the scope of DNA barcoding, in species description or in molecular detection.

Acknowledgement

The authors thank Guido Drovandi, Alessandro Giacomini, Gabriele Giammusso, Giulia Brunori and Federico Russo. This study was partially supported by the FLAGSHIP 'InterOmics' project (PB.P05) funded by the Italian MIUR and CNR institutions.

REFERENCES

Arisi I, D'Onofrio M, Brandi R *et al.* (2011) Gene expression biomarkers in the brain of a mouse model for Alzheimer's Disease: mining of microarray data by logic classification and feature selection. *Journal of Alzheimer's Disease*, **24**, 721–738.

Bertolazzi P, Felici G, Weitschek E (2009) Learning to classify species with Barcodes. *BMC Bioinformatics*, **10**, 1–12.

Bertolazzi P, Felici G, Lancia G (2010) Application of feature selection and classification to computational molecular biology. *Biological Data Mining*(eds Chen JK & Lonardi S), pp 257–294 Chapman & Hall, FL, USA.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Science of the United States of America*, **106**:12794–12797

Felici G, Truemper K (2002) *A MINSAT approach for learning in logic domains*. *Informs Journal on Computing*, **14**, 20–36.

Felici G, Truemper K (2006) The lsquare system for mining logic data. *Encyclopedia of Data Warehousing and Mining*, (eds Wang J) Idea Group Reference, **2**: 693–697.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, **183**, 63–98.

Ratnasingham S, Hebert PDN (2007) Bold: the Barcode of Life Data System. *Molecular Ecology Notes*, **7**, 355–364.

Sarkar IN, Trizna M (2011) The barcode of life data portal: bridging the biodiversity informatics divide for DNA barcoding. *PLoS ONE*, **6**, e14689.

Schoch CL, Seifert KA, Huhndorf S *et al*., and Fungal Barcoding Consortium (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Science of the United States of America*, **109**, 6241–6246.

Tan P, Steinbach M, Kumar V (2005) *Introduction to Data Mining*. Addison Wesley, MA, USA.

van Velzen R, Weitschek E, Felici G, Bakker FT (2012) DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE*, **7**, e30490.

Weitschek E, van Velzen R, Felici G (2011) Species classification using DNA Barcode sequences: A comparative analysis. IASI-CNR Report 11-07

Weitschek E, Lo Presti A, Drovandi G *et al*. (2012a) Human polyomaviruses identification by logic mining techniques. *BMC Virology Journal*, **9**, 58.

Weitschek E, Felici G, Bertolazzi P (2012b) MALA: a microarray clustering and classification software. *DEXA Workshops*, **2012**, 201–205.

P.B. and G.F. designed research. E.W. and R.v.V wrote the manuscript. All other authors helped to draft and review the manuscript. E.W. and G.F. designed and developed the BLOG 2.0 software. R.v.V. suggested improvements for BLOG. E.W. and R.v.V. conceived and performed the experimentations. All authors read and approved the final manuscript.

Data Accessibility

The Blog 2.0 software system, the user manuals and sample data sets are available on http://dmb.iasi.cnr.it/blog-downloads.php in its various versions.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

BLOG 2.0 offline user interface manual