Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Evolutionary Bioinformatics

EASER: Ensembl Easy Sequence Retriever

Emanuel Maldonado[1], Imran Khan[1,2], Siby Philip[1,2], Vítor Vasconcelos[1,2] and Agostinho Antunes[1,2]

[1]CIIMAR/CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal. [2]Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto, Portugal.

ABSTRACT: The rapid advances in genome sequencing technologies have increased the pace at which biological sequence databases are becoming available to the broad scientific community. Thus, obtaining and preparing an appropriate sequence dataset is a crucial first step for all types of genomic analyses. Here, we present a script that can widely facilitate the easy, fast, and effortless downloading and preparation of a proper biological sequence dataset for various genomics studies. This script retrieves Ensembl defined genomic features, associated with a given Ensembl identifier. Coding (CDS) and genomic sequences can be easily retrieved based on a selected relationship from a set of relationship types, either considering all available organisms or a user specified subset of organisms. The script is very user-friendly and by default starts with an interactive mode if no command-line options are specified.

KEYWORDS: sequence analysis, bioinformatics, molecular evolution, genomics, data curation, databases

CITATION: Maldonado et al. EASER: Ensembl Easy Sequence Retriever. *Evolutionary Bioinformatics*. 2013:9 487–490 doi: 10.4137/EBO.S11335.

TYPE: Technical Advance

FUNDING: The authors acknowledge the Portuguese Foundation for Science and Technology (FCT) for financial support to IK (SFRH/BD/48518/2008) and SP (SFRH/BD/47938/2008). This work was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Program and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610), PTDC/AAC-CLI/116122/2009 (FCOMP-01-0124-FEDER-014029) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490) to AA.

COMPETING INTERESTS: Author(s) disclose no potential conflicts of interest.

CORRESPONDENCE: aantunes@ciimar.up.pt

Functionality

Genomics studies often start with the organization of a dataset retrieved from public databases, such as the Ensembl.[1] The Ensembl database is currently in release 71 (April 2013) and covers genomes from vertebrates and other eukaryotic species, and therefore holding a large amount of genomic data. Ensembl provides well characterized gene and gene family annotations together with well defined gene homologous (orthologs and paralogs) relationships, which are very accurate and a rich source of information for evolutionary and comparative genomics. The Ensembl database is continuously expanding with new genomes being added gradually. It is very important, therefore, to develop faster, easier, and more user friendly methods in order to make them available to the broad community of biologists enabling this important resource to be exhaustively used. At present, the process of creating a dataset is very tedious and time consuming, particularly if the user must search and download every sequence individually, making it almost impossible to use such important and valuable resource. To overcome such hurdles, users currently have a few available tools for downloading data from Ensembl, namely BioMart,[2] the Application Programming Interfaces (APIs), and the retrieve-ensembl-seq[3] program.

The BioMart application, however, can be limiting considering the number of species that can be obtained per query, which is at most five.[3] The APIs are not useful to the common biologist, as high programming skills are required to maximize its use. Although the retrieve-ensembl-seq program is more user-friendly, its dataset preparation is cumbersome with some limitations regarding the taxon level, where no option exists to choose the desired species, an option very much needed for comparative genomics. In addition, users have to select multiple options from web-based non-intuitive menus making the overall process tedious, particularly when building a large dataset, thus diminishing the overall purpose of fast and large-scale data preparation.

In order to address this problem, we have developed a much more user friendly application for biologists, which has the advantage of allowing all types of genomics features to be downloaded (Table 1). This includes complete genes, coding-sequences (CDS), peptide, exon, intron, 5' and 3' UTR, upstream, downstream, as well as both extended regions, 5' end and 3' end, as per user defined lengths in base-pairs (bp) based on a given relationship (Table 2) for a given gene ID. Hence, it becomes more efficient to download Ensembl defined genomic features for as many gene IDs and as many species as per user requirement. Following the download, the user often needs to format the descriptions of sequences, such as the word limits and the use of special characters[4] for the requirements of the downstream analyses. We further provided the user with the option of naming the downloaded sequences in one of the given four EASER-specific options: 1) species name with gene symbol; 2) Ensembl ID; 3) Ensembl ID with species name and gene symbol ; and 4) species name abbreviated and gene symbol (Fig. S2), thus simplifying the user job of data downloading and formatting.

Here we present EASER (easer.py), a simple Python (http://www.python.org/) program for the retrieval of large amounts of sequence data from the Ensembl database. In the case of coding-sequences, the termination codon is removed. The script provides an easy, fast, and effortless way for downloading the homologous sequences from any given number of available species selected by the user. The user can select all the desired options at once or one by one from interactive mode, and in a matter of seconds or a few minutes the sequences are saved in the user personal computer in FASTA file format. The Perl (http://www.perl.org/) script enamer.pl can be used independently for renaming the NCBI and Ensembl specific sequence descriptions in any selected format from the available options provided interactively.

Our application provides the user the access to an alternative and oversimplified way of working, focused in the personal computer environment. The PyCogent[5] library implements

Table 1. Available sequence feature types and corresponding options.

SEQUENCE FEATURES	OPTION	SYNTAX	NOTES
Genomic	g	-d g	Complete gene
Coding (CDS)	c	-d c	Sequence feature used by default. Without stop codon
Peptide	p	-d p	–
Exon	e	-d e	These are numbered in their order from Ensembl
Intron	i	-d i	These are numbered in their order from Ensembl
UTR 5'	u5	-d u5	–
UTR 3'	u3	-d u3	–
UTR 5' and 3'	u53	-d u53	Both 5' and 3' UTRs
Flanking 5' (upstream)	f5;YSIZE	-d f5;YSIZE	YSIZE is any positive integer chosen by the user for the length of extended upstream region in bp. Eg, -d f5;750
Flanking 3' (downstream)	f3;ZSIZE	-d f3;ZSIZE	ZSIZE is any positive integer chosen by the user for the length of the extended downstream region in bp Eg, -d f3;650
Flanking 5' and 3' (up and downstream)	f53; YSIZE; ZSIZE	-d f53; YSIZE; ZSIZE	Both upstream and downstream. YSIZE and ZSIZE are any positive integer chosen by the user. YSIZE for upstream region extension and ZSIZE for downstream region extension; both in bp Eg, -d f53;750;650

Table 2. Ensembl defined genomic feature relationships and corresponding options.

RELATIONSHIP	OPTION	SYNTAX	NOTES
Orthologs			
apparent_ortholog_one2one	0	-R 0	Single gene from each species, related to the duplication node
ortholog_one2one	4	-R 4	Depending on the number of genes found in each species. Default option
ortholog_one2many	3	-R 3	Depending on the number of genes found in each species
ortholog_many2many	2	-R 2	Depending on the number of genes found in each species
possible_ortholog	6	-R 6	When the duplication have species-intersection-score ≤ 0.25
Paralogs			
within_species_paralog	10	-R 10	Relation between two genes of the same species with ancestor duplication node.
other_paralog	5	-R 5	Related as member of a broader "super-family"
Projection			
projection_altered	7	-R 7	Gene with one or more novel transcripts, with a known gene from Human or Mouse as ortholog
projection_unchanged	8	-R 8	Gene with one or more novel transcripts, with a known gene from Human or Mouse as ortholog
Gene Split			
contiguous_gene_split	1	-R 1	Little or no overlap between the gene fragments present in same strand close to each other (<1MB)
putative_gene_split	9	-R 9	Little or no overlap between the gene fragments present in different sequence regions in the assembly.

access to Compara API of the Ensembl database. The use of this API ensures that a constant up-to-date access to sequences is gained. The next sections introduce the EASER's options.

Options

Command-line mode. Quick start: Just typing (easer.py –s ENSG00000108511 –a 1) will help the user to download the coding sequences (orthologs one-to-one) for all species available.

To view a list of available options, the user types the option -h (easer.py -h) in the terminal (Fig. S1). The options listed enable the user to fulfill different requirements. For instance, option -s enables the user to specify a single Ensembl ID (-s ENSG00000108511) or an input file containing multiple Ensembl IDs (-s ensemblids.txt) organized in one single column for a set of genes. Other options include option -R for the Ensembl homology relationship choice (see Table 2), option -r for the Ensembl database release number, option -a for the species selection (if omitted, the user will be prompted for the species selection), option -o for output file name (by default Ensembl IDs are used as file names), and option -c for sequences renaming.

Depending on the user requirement, option -a can be used to select all species (-a 1) or select a group of species specified in a file and provided by the user (-a mammals.txt). In this case, the possibility to provide a file with the required group of species is given to the user (e.g., for mammals group, see file mammals.txt provided with EASER archive). Similarly, it is possible to create a file with desired species from the complete species list (see Interactive Mode). Therefore, the user has freedom to choose any species or taxon regarding the diverse research interests.

By default the script downloads coding-sequences, thus the user must specify option -d in order to obtain the desired feature (see Table 1). For example, by typing the single command (easer.py –s ENSG00000108511 -a 1 -r 67 -o mydata.fas -d e) all one to one orthologs for exons from all the available species in the release 67 of the Ensembl database will download for the given Ensembl gene ID, and save in mydata.fas output file. If multiple Ensembl IDs are specified in a file (e.g., -s ensemblids.txt), the results are saved separately in different output files where the naming is done in accordance to the output file name given and the position of the Ensembl ID in input file following its top-down order (e.g., mydata1.fas is meant for download results from first Ensembl ID in ensemblids.txt).

Since the PyCogent[5] library (currently in version 1.5.3) does not provide automated updates of the organisms list, our script fulfills this gap by providing an option (-m) which prompts for the species scientific and common names to be added to the list, thus adding the new species to the initial species list.

Interactive mode. In the case where no options are given in the command-line, the easer.py script enters the interactive mode and a set of successive questions is posed to the user in order to obtain the necessary options for data retrieval. When the option -a is not selected in command-line, the script enters this mode to present a list of available species, from which the user can select their option by choosing the corresponding species numbers separated by commas. The script will then fetch all the data.

Results

We provide the user an easy access to the retrieval of biological sequence features from the Ensembl database for comparative genomics and evolutionary studies. The desired genomic features (Table 1 and 2) can be easily downloaded for the corresponding Ensembl ID and the multi-sequence file can be prepared given the user requirements. This set of sequences can further be renamed and readily used for downstream analyses, thereby saving lots of time and energy. Our script is a valid tool that interacts with well-defined Ensembl features and an easy-to-use alternative to the currently available options. Since it runs on the user's personal computer, the EASER program is a more effective and practical way of generating a sequences dataset from the Ensembl database directly to a FASTA file.

Given the EASER is command-line driven, the efficiency is greatly improved in building appropriate datasets. If the user has multiple Ensembl IDs to use as query, these can be provided in a file and thus avoid running the script several times for the same options. The user can even start several instances of this script simultaneously in several terminal windows, choosing options wisely for every case and make the download of data instantly for every Ensembl ID or set of IDs.

Our script has been successfully used by us and is cited in several manuscripts under review and published from our group.[6]

EASER is open for further improvements, which can be performed in the near future regarding the user's needs.

Availability and requirements

EASER is currently in version 1.7.0 and is freely available under GNU General Public License upon request or on the web at http://easer.sourceforge.net/. This script is implemented in Python (http://www.python.org) and is based in the PyCogent[5] library, for which installation instructions can be found in Quick Installation at http://www.pycogent.org. Presently PyCogent[5] library supports Linux, Windows 64-bit and MacOS, enabling the use of EASER in any of these systems.

EASER was developed for UNIX/Linux environment. In order to start using the script, and provided that 1) Python and 2) PyCogent[5] library are already installed in the user's system, the user must give it execution permissions and it will be ready to run from its current location directory. The user can also place the scripts in a binaries directory, enabling the use of EASER from any working directory.

Acknowledgements

We are thankful to the anonymous reviewers for their valuable comments and suggestions which lead to the improvement of the manuscript.

Author Contributions

Conceived and designed the experiments: EM, IK, SP, AA. Analyzed the data: EM, IK, SP. Wrote the first draft of the manuscript: EM. Contributed to the writing of the manuscript: EM, IK, SP, AA. Agree with manuscript results and conclusions: EM, IK, SP, VV, AA. Jointly developed the structure and arguments for the paper: EM, AA. Made critical revisions and approved final version: EM, IK, SP, AA. All authors reviewed and approved of the final manuscript.

DISCLOSURES AND ETHICS

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

Supplementary Data

Supplementary Figure 1. Help menu for easer.py script. Type easer.py -h in terminal to access this menu.

Supplementary Figure 2. Flow chart exhibiting the functionality of EASER.

REFERENCES

1. Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. *Nucleic Acids Res*. 2012;40(Database issue):D84–90.
2. Kasprzyk A, Keefe1 D, Smedley D, et al. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*. 2004;14(1):160–9.
3. Sand O, Thomas-Chollier M, van Helden J. Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl. *Bioinformatics*. 2009;25(20):2739–40.
4. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586–91.
5. Knight R, Maxwell P, Birmingham A, et al. PyCogent: a toolkit for making sense from sequence. *Genome Biol*. 2007;8(8):R171.
6. Philip S, Machado JP, Maldonado E, et al. Fish Lateral Line Innovation: insights into the evolutionary genomic dynamics of a unique mechanosensory organ. *Mol Biol Evol*. 2012;29(12):3887–98.