BMC Bioinformatics

BioMed Central

Software

Open Access

eL-DASionator: an LDAS upload file generator

Vincent Negre and Christoph Grunau*

Address: Institut de Génétique Humaine, CNRS UPR 1142, 141 rue de la Cardonille, 34396 Montpellier, France

Email: Vincent Negre - developer@methdb.net; Christoph Grunau* - cgrunau@igh.cnrs.fr

* Corresponding author

Published: 07 May 2004

BMC Bioinformatics 2004, **5**:55

Received: 29 January 2004
Accepted: 07 May 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/55

Abstract

Background: The Distributed Annotation System (DAS) allows merging of DNA sequence annotations from multiple sources and provides a single annotation view. A straightforward way to establish a DAS annotation server is to use the "Lightweight DAS" server (LDAS). Onto this type of server, annotations can be uploaded as flat text files in a defined format. The popular Ensembl ContigView uses the same format for the transient upload and display of user data.

Results: In order to easily generate LDAS upload files we developed a software tool that is accessible via a web-interface http://atgc.lirmm.fr/eldasionator.html. Users can submit their DNA sequences of interest. Our program (i) aligns these sequences to the reference sequences of Ensembl, (ii) determines start and end positions of each sequence on the reference sequence, and (iii) generates a formatted annotation file. This file can be used to load any LDAS annotation server or it can be uploaded to the Ensembl ContigView.

Conclusion: The eL-DASionator is an on-line tool that is intended for life-science researchers with little bioinformatics background. It conveniently generates LDAS upload files, and makes it possible to generate annotations in a standard format that permits comfortable sharing of this data.

Background

DNA sequences are carriers of information. This information has to be elucidated by experiments and can then be associated with the corresponding regions of the genome. Biological research is providing us with a multitude of such annotations to the genomic DNA sequence. However, it is difficult to compare and combine this data and to get a global overview. In order to make these heterogeneous sequence annotations readily accessible, Lincoln Stein and colleagues developed an open standard for biological data exchange called the Distributed Annotation System (DAS) [1,2]. DAS consists of three parts: (i) a central reference server with the genomic DNA sequence, (ii) various annotation servers with specialised annotations and (iii) a client viewer which displays sequence information from the reference server and various annotations layers depending on what the user wants to include. A popular reference server is maintained by the Ensembl project [3,4]. The associated client viewer is the "ContigView" display of the Ensembl genome browser. The Lightweight Distributed Annotation System (LDAS) package provides a convenient way to install an annotation server [5]. It is based on the MySQL database system, the Apache web server, the Perl programming language, and the Bioperl module library [6]. Perl scripts, provided with the LDAS distribution, allow loading the annotation server with tab-delimited text files. These files consist of two sections: annotation and reference. The annotation section contains the annotations for the reference sequence. The reference section is a listing of the reference sequences for which annotations are provided in the annotation section. Naturally, annotations must be correctly located on the reference sequence where they belong to. This positioning could be done by manual alignment of the annotated sequence with the reference sequence, and the editing of a corresponding text file. However, such a procedure is error prone and labour intensive, in particular since it has to be repeated each time reference and/or annotated sequences are updated. For a small number of FASTA formatted sequences the Ensembl project provides a service that places the sequences into the genomic context and generates a display [7]. But only up to 30 sequences can be treated each time with this tool. Users having more sequences to deal with will probably generate their own software, provided they possess the necessary computer skills. However, for a large number of biologists who do not have such bioinformatics resources at their disposal, the generation of annotation files on a regular basis will not be possible. This is a handicap for a broader use of DAS. Therefore, we sought to create an easy to use software tool that is capable of aligning large numbers of annotated sequences with the reference sequences, determine start and end positions of each annotated sequence, and generate a formatted file ("load file") that can be uploaded temporarily to the Ensembl ContigView or can be used to load dedicated LDAS annotation servers.

Implementation

Our tool, the eL-DASionator, consists of several consecutively called Perl-CGI scripts with web-interfaces. In order to align the annotated sequence with the reference sequences, we use BLAST [8] with default parameters. The alignment is performed locally on our server. Reference sequences are regularly updated from Ensembl via FTP [9]. We provide genomic reference sequences for all species available in the Ensembl project.

The generation of a LDAS upload file with the eL-DASionator consists of three major steps. First, the sequences for which the user has annotations available are submitted via a web form. Sequences must be DNA in concatenated FASTA format. The user chooses class, type and subtype of annotations. These entities are basic annotation terms like "gene" and "exon". Submitted sequences are aligned to the reference sequence using BLAST. Sequences of less than 7 nucleotides are discarded because they produce no significant alignment. Since the BLAST algorithm is a local alignment algorithm, it fails occasionally to align the entire sequence. Therefore, we provide two alternatives to determine the location of the annotated sequence on a genome. If the user selects the option "calculated borders" on the initial submission form, the start and stop positions correspond to the first and last bases of the submitted sequence. If the user selects "alignment borders", the start and end positions correspond to the first and last nucleotides of the BLAST alignment. In this case, the alignment length can be shorter than the annotated sequence.

As the second step, the alignments are displayed on a webpage, and the user is requested to verify the alignments. In order to make manual verification more convenient, results are split into groups of 200 annotated sequences. The user can select up to two reference sequences per annotated sequence (e.g. overlapping contigs), or he can reject the alignment by selecting "no reference sequence". After hitting the submit button, the LDAS load file is generated automatically. The file content is displayed on a web form where the user can edit the file if he desires. Finally, after the user has confirmed the annotations, an automatic verification checks whether tabulation and column numbers for each section are correct. Then, the user can save the load file, which is now ready for use.

Results

Initially, we used the eL-DASionator to load our own LDAS server, which provides annotations about DNA methylation and is linked to the DNA methylation database MethDB [10,11]. This database contains currently methylation data for 121 different sequences. In order to confirm that our tool is able to handle large data sets, we treated the 31312 sequences generated by the CpG-island tagging project [12]. Also in this case, eL-DASionator successfully generated LDAS load files. Just about half a day on-screen editing was necessary to verify the alignments. For the moment, our tool uses only the reference sequences of the Ensembl server. This is one of the most popular projects and it provides reference sequences for man and for most model organisms.

Conclusion

DAS is a powerful system to share specialised sequence annotations and to combine the expertise of many laboratories into a common resource. Its main drawback is the need for keeping the positions of annotated sequences in correspondence with the permanently changing reference sequence. Manual editing of the files is too laborious and presents a major limit to the further success of the DAS. The eL-DASionator provides a solution to this problem and reduces the effort to share biological information. Thanks to our tool, users with little bioinformatics background can now easily generate annotations according to the standardized DAS/1 specification by either sending load files generated by the eL-DASionator directly to the Ensembl ContigView or by loading remote LDAS annotation servers with them. The eL-DASionator source code can be obtained from the project home page. Installation instructions are supplied as README file [see 1].

Availability and requirements

Project name: eL-DASionator

Project home page: http://atgc.lirmm.fr/eldasionator.html

Operating system(s): server – Linux, client – platform independent

Programming language: Perl

Other requirements: Apache, Bioperl

License: General Public License

Any restrictions to use by non-academics: none

Authors' contributions

VN carried out the main programming work; CG conceived the software tool and participated in the programming.

Additional material

Additional File 1

contains instructions about how to install the eL-DASionator on a local server using the archive "eldasionator_source.tar.gz". This archive can be downloaded from the eL-DASionator home page.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-55-S1.txt]

Acknowledgements

This work was supported by the Génopole Languedoc-Roussilon and by a grant of the European Commission (HPRN-CT-2000-00089). We are grateful for technical support from the bioinformatics group of the Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM).

References

1. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2:**7.
2. **biodas.org** [http://www.biodas.org/]
3. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.
4. **Ensembl Genome Browser** [http://www.ensembl.org/]
5. **The Lightweight Distributed Annotation Server (LDAS)** [http://biodas.org/servers/LDAS.html]
6. **The Bioperl Project** [http://bioperl.org/]
7. **BLASTView** [http://www.ensembl.org/Multi/blastview]
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
9. **Ensembl anonymous FTP site** [ftp://ftp.ensembl.org/]
10. Amoreira C, Hindermann W, Grunau C: **An improved version of the DNA Methylation database (MethDB).** *Nucleic Acids Res* 2003, **31:**75-77.
11. **DNA methylation database** [http://www.methdb.net/]
12. **CpG island tagging project** [http://www.sanger.ac.uk/HGP/cgi.shtml]