Published online 21 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W161–W164***
doi:10.1093/nar/gkq426

SOBA: sequence ontology bioinformatics analysis

Barry Moore, Guozhen Fan and Karen Eilbeck*

Department of Human Genetics, University of Utah, Salt Lake City, Utah, UT 84112, USA

Received February 4, 2010; Revised April 30, 2010; Accepted May 6, 2010

ABSTRACT

The advent of cheaper, faster sequencing technologies has pushed the task of sequence annotation from the exclusive domain of large-scale multi-national sequencing projects to that of research laboratories and small consortia. The bioinformatics burden placed on these laboratories, some with very little programming experience can be daunting. Fortunately, there exist software libraries and pipelines designed with these groups in mind, to ease the transition from an assembled genome to an annotated and accessible genome resource. We have developed the Sequence Ontology Bioinformatics Analysis (SOBA) tool to provide a simple statistical and graphical summary of an annotated genome. We envisage its use during annotation jamborees, genome comparison and for use by developers for rapid feedback during annotation software development and testing. SOBA also provides annotation consistency feedback to ensure correct use of terminology within annotations, and guides users to add new terms to the Sequence Ontology when required. SOBA is available at http://www.sequenceontology.org/cgi-bin/soba.cgi.

INTRODUCTION

Genome annotation

A fully sequenced and assembled genome is only the first step towards understanding the information encapsulated in the genome sequence. Genome annotation is the process of layering biologically relevant knowledge upon a sequenced and assembled genome. Annotation is the key to making use of the genome in downstream analyses, and the quality of annotation will make or break these experiments. The annotation process involves compiling a wide range of experimental and computational evidence such as alignments to EST and cDNA libraries, protein databases and repeat libraries from the same or similar organisms, gene predictions from either *ab initio* or evidence-based gene prediction algorithms; and finally, fully descriptive gene models created synthesizing all available evidence.

Historically, the large model organism databases such as FlyBase (1) and WormBase (2) used bioinformatics analysis backed by teams of curators to interpret the various sources of evidence and annotate the gene models. Recently, with the advent of faster, cheaper sequencing, it is the downstream analysis and data-handling that has become the bottle-neck. These post-sequencing tasks may surpass the data production in cost (3). Of the 1377 eukaryotic genomes sequence projects listed in the GenomesOnLine Database version 3.O (4) only 169 are marked ‘complete and published’ whereby the data is available via public genome archives.

There are many well established genome annotation pipelines serving the large sequencing and annotation centers such as the Ensembl (5) pipeline. However, with decreasing cost and increasing speed of sequencing, whole genome sequence annotation is entering the scope of single laboratories and small consortia of biologists. Locally installable, automated annotation pipelines such as GenDB (6), Gramene pipeline (7) and MAKER (8) are being utilized to make the feature calls and produce the annotations for many newly sequenced and assembled genomes, such as that of the Planarian: *Schmidtea mediterranea* (9).

Although the genome sequences are usually stored and maintained in relational databases such as Chado (10), the main currency of annotation has become the tab delimited flat file, which can be easily shared between researchers and used as the substrate for visualization and analysis programs. The flat file format GFF (http://www.sanger.ac.uk/resources/software/gff/spec.html) emerged during the Human Genome Project and several varieties have since proliferated. This proliferation caused problems as the formats may look similar but often, different groups have either used different terms to mean the same thing or the same term has slightly different meanings. This is problematic for groups parsing and analyzing data from multiple sources. The Sequence Ontology (SO) (11) has brought standardization to terminology and semantics captured by these flat file formats by categorizing the terms used to describe sequence features into an ontology. This formalization is used to name the

*To whom correspondence should be addressed. Tel: 801 585 9934; Fax: 801 581 7796; Email: keilbeck@genetics.utah.edu

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

features in the Generic Model Organism Database (GMOD) group's (http://www.gmod.org) revision of the format, known as GFF3. GFF3 (http://www.sequenceontology.org/resources/gff3.html) is commonly used as the input and output of GMOD tools as well as the release format for many model and emerging model organism databases.

Using the SO to characterize the type of feature and the relations between features has unified the vocabulary used by the community. The ontology also provides the ability to specify the feature at the deepest level known but query the data at different levels of specificity. It provides an abstraction between the data and the software that handles the data.

There are many examples that illustrate the utility of the GFF3 format. Newly created annotations, either made by manual annotation for example using Apollo (12), or automated annotation pipelines such as MAKER export the annotation in GFF3. It is therefore natural that many model organism databases also release their sequences in this format such as DictyBase (http://dictybase.org/Downloads/), the database of *Dictyostelium discoideum* and WormBase (ftp://ftp.wormbase.org/pub/wormbase/datasets), the database of *Caenorhabditis* species. Recently with the advent of whole genome sequencing, the variant files produced by endeavors such as the 1000 genomes project are also structured to meet the standard of GFF3 such as the variant calling format (http://www.1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:variant_call_format). Software for visualization and analysis of annotation are consumers of GFF3 such as Gbrowse (13) and Comparative Genomics Library (14).

Here, we provide a tool to perform analysis over newly created genome annotations, specified in the GFF3 format. We are addressing four main use cases.

(i) Analysis for emerging model system groups: SOBA provides a first set of statistics to summarize a newly sequenced and automatically annotated genome.
(ii) Comparative genomics: SOBA provides an overview of the structure of genome annotations between multiple species.
(iii) Analysis for developers producing tools that produce a genome annotation: SOBA provides a rapid set of statistics with which to evaluate the performance of a tool such as a gene finder.
(iv) Promote annotation consistency: SOBA allows users to find annotation inconsistency in their files, with regards to ontology usage, and provides several steps to fix the problems.

USING SOBA

Input

The input to SOBA is a genome annotation comprising of one or more files in the GFF3 format. The files may be uploaded either from a local directory, or via a URL. The upper limit on total file size is 1.5 GB, which corresponds roughly to 12 million sequence features. A GFF3 file is tab-delimited to nine columns, which capture the details of each feature such as its source (the program or resource that called the feature), its start and end coordinates relative to a given landmark such as a contig or chromosome and its SO type. A sample of a file is shown in Figure 1A.

Calculations

For each data source, SOBA provides counts for each kind of sequence feature appearing in column 3 of the GFF3 file. For each of these features, the minimum, maximum, mean, median and footprint of the feature's collective length on the genome is calculated. The footprint of a feature type is defined as the cumulative, non-redundant nucleotide count of all features of a given type, divided by the total nucleotide length of the sequence represented in the file (Figure 1B).

A graphical histogram representing the distribution of feature length is presented for each feature by data source (Figure 1F).

Intron density is a measure of the number of introns per protein, as described by Roy *et al.* (15). It is calculated by dividing the number of coding introns in an mRNA annotation by the length of the encoded protein (Figure 1E).

For each ontology term used, the is a path back to the root node in the ontology is parsed, to produce a representation of both the terms used and their transitive parents (Figure 1D). The terms in the ontology image are clickable and link directly that term in the miSO ontology browser (www.sequenceontology.org/cgi-bin/miso.cgi).

Data presentation and visualization

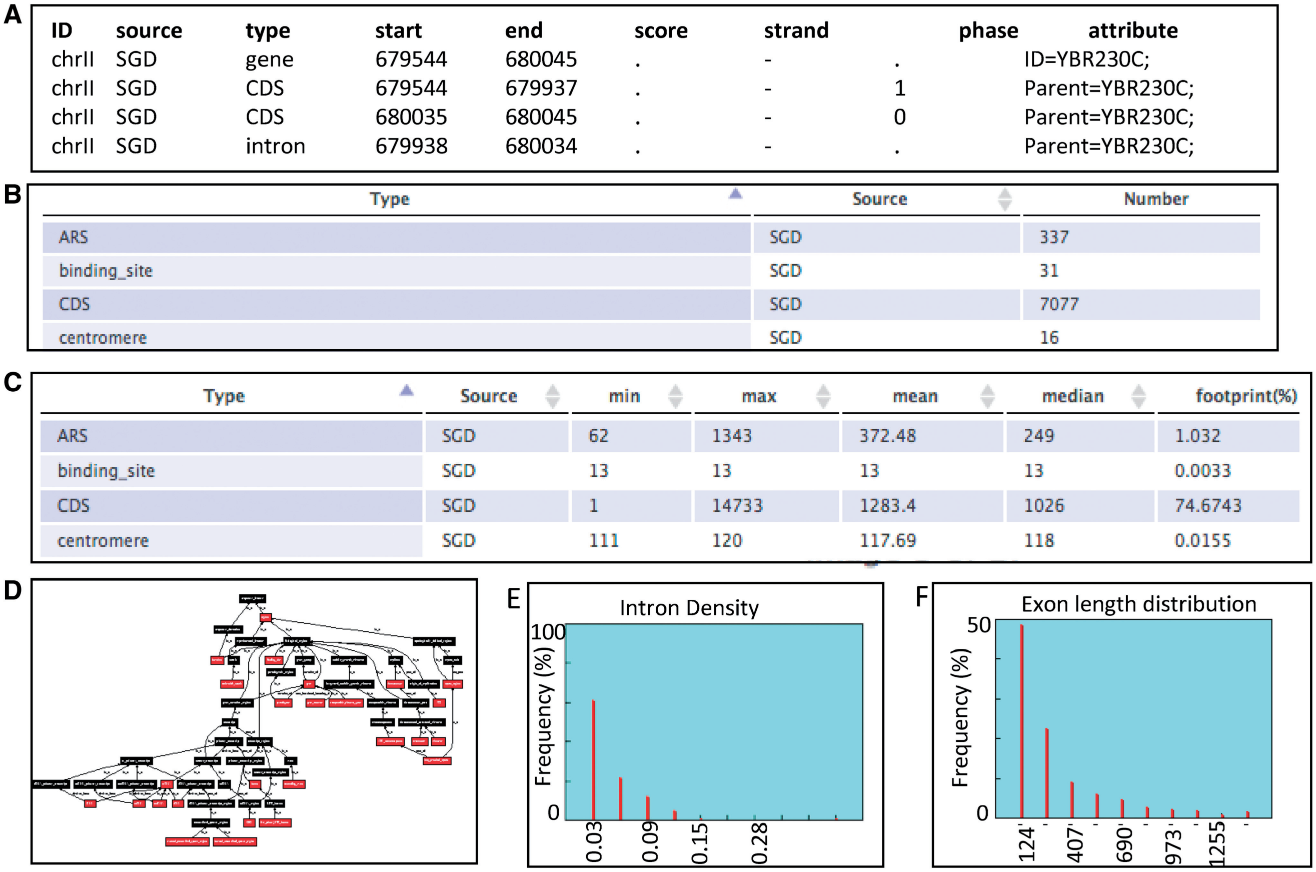
Upon upload, the user must select the features and sources to be displayed. The compact visualization of the results allows the user to browse the results of a query one at a time. In addition, SOBA provides validation of the terminology used in the uploaded file, and suggests corrections for the user. There are three groups of invalid features. The terms used, that are a synonym of a SO term are highlighted and the correct term is shown. Terms that have incorrectly formatted case are also shown, with the correct term. Finally, terms that are not part of the SO are linked to the term request tracker where the user can make a new term suggestion to the ontology developers.

Outputs

The output of SOBA is a web-based summary of the genome annotation for user-selected features and sources of data. In addition to the web-based graphical and tabular output, users may also export the data to their local computer for use in generating images and reports for article and grant preparation. These results may be exported as PDF files, tab-delimited text files, HTML pages and GIF images.

Implementation

SOBA is implemented with maintenance and extensibility as key features. The web server uses the Perl-



A

ID	source	type	start	end	score	strand	phase	attribute
chrII	SGD	gene	679544	680045	.	-	.	ID=YBR230C;
chrII	SGD	CDS	679544	679937	.	-	1	Parent=YBR230C;
chrII	SGD	CDS	680035	680045	.	-	0	Parent=YBR230C;
chrII	SGD	intron	679938	680034	.	-	.	Parent=YBR230C;

B

Type	Source	Number
ARS	SGD	337
binding_site	SGD	31
CDS	SGD	7077
centromere	SGD	16

C

Type	Source	min	max	mean	median	footprint(%)
ARS	SGD	62	1343	372.48	249	1.032
binding_site	SGD	13	13	13	13	0.0033
CDS	SGD	1	14733	1283.4	1026	74.6743
centromere	SGD	111	120	117.69	118	0.0155

Figure 1. The input and output of the SOBA tool. (**A**) Small portion of a GFF3 file, including the column headings. (**B–F**) Screen shots of the output of SOBA. (B) The primary counts for each feature type per data source. (C) The simple statistics for the lengths of each feature including the mean, median and footprint of the feature on the genome. (D) A high-level view of all of the SO terms used in the genome annotation and the transitive *i_sa* relations back to the root node. A large format version of this panel is available at http://sequenceontology.org/resources/images/Figure1D.gif. (E) The distribution of intron density of protein coding genes (number of coding introns/length of polypeptide sequence). (F) An example of a sequence feature length distribution showing the distribution of lengths of annotated exons.

based, Model-View-Controller (MVC) structured CGI::Application as the underlying framework. This framework consolidates the computation and logic of SOBA into Perl modules separate from those implementing the Graphical User Interface (GUI) front-end and presentation of results. Web view and downloadable reports are generated with a collection of templates. The Perl-based Template Toolkit (http://www.template-toolkit.org/) package provides a robust and extremely flexible template engine used for generating these 'Views' providing ease of maintenance and extensibility for the web application.

The SOBA web server utilizes the JQuery JavaScript library (http://jquery.com/) to provide a convenient and intuitive user interface. Various JQuery plugins allow SOBA to present a large amount of information in a manageable way with accordion views of different data types, sortable tables, graphics slide shows and asynchronous page refreshes that users have come to expect of Web 2.0 applications.

The Graphviz (http://www.graphviz.org/) package is utilized to generate graphical views of SO graphs. This provides a valuable overview of the SO terms used in the GFF3 file under analysis, and is presented in the same format as miSO the SO browser. Nodes in the graph view of the GFF3 file a links to the same terms within the SO allowing users to easily view details of the terms and see how terms in their file fit into the larger framework of the SO.

The Perl-based GD modules along with the underlying C-based GD Graphics Library (http://www.boutell.com/gd/) are used to generate charts.

All Perl modules discussed above as well as others used to implement SOBA are available from CPAN (http://www.cpan.org/). SOBA is released under the Artistic License, which allows for modification and redistribution by all users and as such is compatible with the Open Source Initiative's (http://www.opensource.org/) definition of Open Source software.

DISCUSSION AND CONCLUSIONS

The NCBI via database resource (16) provides a statistical summary of the genome assemblies with annotations that they maintain, via Entrez Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/static/MapViewerHelp.html). Although both tools share several statistics, SOBA provides analysis of any GMOD compliant genome annotation, either a hot off the press new sequence or an existing well known genome. SOBA also addresses the semantics of the annotation with the summary of ontology term usage, whereas the NCBI lacks ontological markup of its sequence and therefore does not offer this capability.

SOBA includes on-line documentation via the SO Wiki (http://www.sequenceontology.org/wiki/index.php/SOBA_-_Sequence_Ontology_Bioinformatics_Analysis) and includes both a bug tracker and a feature request tracker for continued development and maintenance of the tool. The MVC architecture of the tool allows for extensibility, and it is envisaged that over time, new tests and views of the genome annotation will be added to meet demand. The use of SOBA may also increase ontology development. When the input file contains terminology not in the ontology, the user is directed to a form page to make a request for a new term.

SOBA was created to provide a simple tool for genome annotation summary that is compliant with the current GMOD tools and pipelines that produce and use genomic information. It is complementary to a genome browser in that it shows an overview of the data and its structure rather than a nucleotide level view of the topological relationships between features. Genome annotation is ultimately an iterative process where groups run and re-run analysis, varying the input and parameters to fine-tune the annotation of their organism. SOBA can quickly provide vital feedback to such groups, helping them evaluate the effects of changes in an annotation pipeline. Towards this goal, uploading data to SOBA is also available as a post genome annotation step via the Maker Web Annotation Service (http://www.yandell-lab.org/software/mwas.html), where it is offered as a complement to viewing the newly created annotations in a genome browser.

ACKNOWLEDGEMENTS

We would like to thank Carson Holt, Hadi Islam, Sofia Robb and Eric Ross for testing SOBA. We also thank Dave Clements, Ram Podicheti and Erick Antezana for their valuable feedback.

FUNDING

Funding for open access charge: National Institutes of Health; National Human Genome Research Institute (Grant HG004341 to K.E.).

Conflict of interest statement. None declared.

REFERENCES

1. Drysdale,R. (2008) FlyBase: a database for the Drosophila research community. *Methods Mol. Biol.*, **420**, 45–59.
2. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
3. Metzker,M.L. Sequencing technologies–the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
4. Liolios,K., Chen,I.M., Mavromatis,K., Tavernarakis,N., Hugenholtz,P., Markowitz,V.M. and Kyrpides,N.C. The genomes on line database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **38**, D346–D354.
5. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
6. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. *et al.* (2003) GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.
7. Liang,C., Jaiswal,P., Hebbard,C., Avraham,S., Buckler,E.S., Casstevens,T., Hurwitz,B., McCouch,S., Ni,J., Pujar,A. *et al.* (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.*, **36**, D947–D953.
8. Cantarel,B.L., Korf,I., Robb,S.M., Parra,G., Ross,E., Moore,B., Holt,C., Sanchez Alvarado,A. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.
9. Robb,S.M., Ross,E. and Sanchez Alvarado,A. (2008) SmedGD: the Schmidtea mediterranea genome database. *Nucleic Acids Res.*, **36**, D599–D606.
10. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
11. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
12. Lewis,S.E., Searle,S.M., Harris,N., Gibson,M., Lyer,V., Richter,J., Wiel,C., Bayraktaroglir,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
13. Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9.9.
14. Yandell,M., Mungall,C.J., Smith,C., Prochnik,S., Kaminker,J., Hartzell,G., Lewis,S. and Rubin,G.M. (2006) Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput. Biol.*, **2**, e15.
15. Roy,S.W., Fedorov,A. and Gilbert,W. (2002) The signal of ancient introns is obscured by intron density and homolog number. *Proc. Natl Acad. Sci. USA*, **99**, 15513–15517.
16. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5–D16.