BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 14 2009, pages 1836–1837
doi:10.1093/bioinformatics/btp314

Genome analysis

Apollo: a community resource for genome annotation editing

Lee Ed[1,*], Harris Nomi[1], Gibson Mark[1], Chetty Raymond[2] and Lewis Suzanna[1]

[1]Berkeley Bioinformatics Open Source Projects, Lawrence Berkeley National Laboratory, Berkeley and [2]The Arabidopsis Information Resource, Carnegie Institution of Washington, Stanford, CA, USA

Received and revised on April 13, 2009; accepted on May 8, 2009

Advance Access publication May 13, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: Apollo is a genome annotation-editing tool with an easy to use graphical interface. It is a component of the GMOD project, with ongoing development driven by the community. Recent additions to the software include support for the generic feature format version 3 (GFF3), continuous transcriptome data, a full Chado database interface, integration with remote services for on-the-fly BLAST and Primer BLAST analyses, graphical interfaces for configuring user preferences and full undo of all edit operations. Apollo's user community continues to grow, including its use as an educational tool for college and high-school students.

Availability: Apollo is a Java application distributed under a free and open source license. Installers for Windows, Linux, Unix, Solaris and Mac OS X are available at http://apollo.berkeleybop.org, and the source code is available from the SourceForge CVS repository at http://gmod.cvs.sourceforge.net/gmod/apollo.

Contact: elee@berkeleybop.org

1 INTRODUCTION

There are numerous computational tools that aim to predict gene and exon positions and other biologically significant sequence features in genomic sequence data. To gauge their relative performance, the EGASP project assessed state-of-the-art computational gene prediction methods on the human genome. Comparing over 28 methods, the best ones could accurately predict one of the transcripts from any given gene for close to 70% of the genes. However, when considering all known transcripts, the accuracy of prediction was only 40–50% (Guigo *et al*., 2006). Similarly in nematodes, the nGASP project found the median gene level sensitivity of the best methods was 78% and their specificity was 42% (Coghlan *et al*., 2008). Perhaps, purely computational methods will eventually be accurate enough to rely on alone, but at present, there is still a strong need for biological expertise to refine such automatically generated sequence annotations, and for an intuitive and flexible way for experts to make these refinements.

Apollo is a graphical tool designed for this purpose. Apollo development started in 2000 to aid in the annotation of the *Drosophila melanogaster* genome (Adams *et al*., 2000), with the first public software release in 2002 (Lewis *et al*., 2002). It continues to be used for Drosophila annotation, as well as for annotating human and hundreds of other organisms. Originally developed as a collaboration between FlyBase-BDGP (Flybase Consortium, 2002) and Ensembl, Apollo is now maintained and developed by the Berkeley Bioinformatics Open Source Projects group.

2 NEWLY ADDED FEATURES

Apollo development is driven by community needs. The core software is stable and provides a full complement of annotation browsing and editing capabilities, while new features continue to be added in response to user requests.

2.1 The generic feature format version 3 support

The generic feature format version 3 (GFF3) (http://www.sequenceontology.org/gff3.shtml) is a popular format for representing genomic features and has been adopted by many organism communities, including *D.melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Escherichia coli, Saccharomyces cerevisiae, Mus musculus* and *Homo sapiens*. One of the biggest benefits of GFF3 is its required use of the Sequence Ontology (Eilbeck *et al*., 2005), an ontology designed for describing biological sequence features. Apollo now offers full support for reading and writing GFF3.

2.2 Continuous data support

Continuous data, such as expression levels across the genome, in the context of other automatically or manually derived sequence annotations, can help improve the accuracy of annotations. For example, viewing continuous expression-level data allows biologists to spot neighboring genomic regions that are co-expressed, which is often an indication of missing exons or genes that should be merged. We recently added the ability to load and display continuous transcriptome data by including support for Affymetrix's Integrated Genome Browser (http://igb.bioviz.org) Signal GRaph (.sgr) and the University of California Santa Cruz's Genome Browser (Karolchik *et al*., 2003) wiggle (.wig) data formats. The user interface includes resizable frames and numerical scores that can be seen by hovering the mouse over a specific region.

2.3 Chado database support

Chado (Mungall *et al*., 2007), the official database schema of the GMOD project (http://www.gmod.org), has been designed to handle complex representations of biological knowledge. We have added full read and write support for Chado. The adapter was designed to be RDBMS-agnostic and does not require a PostgreSQL backend as traditionally used with Chado installations. Users can customize

*To whom correspondence should be addressed.

Apollo to reference their specific Chado databases through a configuration file that allows them to define elements such as controlled vocabularies used and sequence features of interest for retrieval.

2.4 Undo of edit operations

Apollo now permits full undo of all edit operations. This has been a highly requested feature for quite some time and has proven to be very popular. Because of the many side effects that an original edit may have triggered, making undo work for all types of edits is a complex operation.

2.5 Graphical interfaces for configuration

Apollo is highly customizable in terms of the data types it can read and how it displays them. In the past users needed to edit very large, complicated text files to adjust these parameters. To make it simpler for users, we have added new graphical interfaces for setting the wide range of options offered by Apollo.

2.6 Remote analysis support

BLAST (Altschul *et al*., 1990) is a commonly used sequence analysis tool. However, users may not have large computational resources available onsite or may prefer not to install the tool and its databases, which can be quite large. Thus, we have added support for communication with the BLAST and Primer BLAST (primer identification) services at the National Center for Biotechnology Information. Apollo's BLAST interface supports blastn, blastx and tblastx, and allows users to customize searching and post-processing criteria for filtering results. Since these analyses can take a while to run, especially if the load on the NCBI servers is high, all requests are handled in the background, allowing users to continue browsing and editing in Apollo as normal. The BLAST results are automatically incorporated into the user's session once the analyses are complete.

3 APOLLO AS AN EDUCATIONAL TOOL

As a result of Apollo's stability, complete set of features and ease of installation and use, adoption has expanded beyond the lab and into the classroom. It is currently used in several universities and high schools as a teaching tool to illustrate the annotation process through a hands-on approach. These institutions include Hamilton College, San Francisco State University, the Dolan DNA Learning Center of Cold Spring Harbor Laboratory and the Science Education Alliance at Howard Hughes Medical Institute.

4 FUTURE DEVELOPMENT

Development on Apollo is an ongoing process, with new data sources, analysis tools, annotation editing approaches and biologically relevant annotation types continually being reviewed for possible inclusion. From these, we are currently developing a specialized editing interface that enables users to modify gene models in direct reference to multiple alignment data. With this interface, users can modify exon boundaries in gene models with respect to multiple sequence alignments to nucleotide and protein data. This provides an efficient means to annotate genes using cross-species data, essential for newly sequenced genomes that lack extensive cDNA collections.

With the growth of large community annotation projects, an easy to use distributed solution is needed. We have started addressing this issue by integrating Apollo with GBrowse (Stein *et al*., 2002), the most popular genome viewer in the GMOD toolkit. This work ties Chado, GBrowse and Apollo together, allowing customized Apollo instances to be launched through Web Start from GBrowse, with both sharing a common Chado backend.

We are planning to take this functionality a step further, by fully integrating the newly designed, AJAX-driven interactive GBrowse (M.Skinner *et al*., manuscript in preparation) with Apollo. One of the benefits of this solution is that rather than having to deal with two separate applications (GBrowse in the web browser and Apollo as a Java application), users will use a common web interface to both view and edit genomic annotations. Since, both applications will be fully contained within the web browser, it will greatly improve the ease of use and user experience for community annotation, as well as making Apollo available to a wider audience.

ACKNOWLEDGEMENTS

We would like to acknowledge and thank Jonathan Crabtree for his contributions to the Chado adapter, Chris Mungall for helping with issues with Chado, Olivier Arnaiz for testing and bug fixes, Scott Cain and David Clements for GMOD support, help and documentation and the Apollo community for the continued support, user requests and help in finding bugs. We would also like to acknowledge and thank Christopher Smith, Bruce Nash, Wei-Jen Chang, Tuajuanda Jordan and Matthew Conte for information on educational programs they have set up that make use of Apollo in a classroom setting.

Funding: National Institute of General Medical Sciences (NIGMS) (National Institutes of Health grant 1R01GM080203-01).

Conflict of Interest: none declared.

REFERENCES

Adams,M.D. *et al*. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Coghlan,A. *et al*. (2008) nGASP—the nematode genome annotation assessment project. *BMC Bioinformatics*, **9**, 549.

Eilbeck,K. *et al*. (2005) The Sequence Ontology: a tool for unification of genome annotations. *Genome Biol*., **6**, R44.

FlyBase Consortium (2002) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*., **30**, 106–108.

Guigo,R. *et al*. (2006) EGASP: the human ENCODE genome annotation assessment project. *Genome Biol*., **7** (**Suppl1**), 1–31.

Karolchik,D. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.

Lewis,S.E. *et al*. (2002) Apollo: a sequence annotation editor. *Genome Biol*., **3**, RESEARCH0082.

Mungall,C.J. *et al*. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*. **23**, i337–i346.

Stein,L.D. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.