WebApolllo:
A WEB-BASED SEQUENCE ANNOTATION EDITOR FOR COMMUNITY ANNOTATION

Instructors:

- Ed Lee[1], elee [at] berkeleybop [dot] org
- Gregg Helt[1], gregghelt [at] gmail [dot] com
- Monica Munoz-Torres[2], monica [dot] cecilia [at] gmail [dot] com

[1] Bioinformatics Open-source Projects, Lawrence Berkeley National Labs

[2] Elsik Lab, Georgetown University

Objectives:

During this demonstration workshop you will:

- Become familiar with the environment and functionality of the WebApollo annotation tool.
- Learn how to corroborate and/or modify automatically annotated gene models using available evidence with WebApollo.

By the end of the demonstration workshop you will be expected to:

- Understand the utility and relevance of WebApollo functions available for the process of manual annotation.
- Understand the importance of verifying the integrity of a gene model, and the relevance of manual annotation (in the context of fundamental and evolutionary biology).
- Be able to use WebApollo for curation of automatically generated gene models.

WebApollo & The Importance of Manual Annotation

In the context of genomics, the process of curation has two phases: i) creation of gene models (which can be as basic as calling open reading frames), and ii) annotation of these gene models using standardized terminology that describes their function, expression patterns, and metabolic network membership. A gene model can be created automatically using cDNA-based evidence such as ESTs, or it can be generated de-novo by looking at local sequence complexity and implementing a mathematical model. This model is often trained using other known gene models called from a 'golden-set' or ‘gold standard’, but it can also run untrained (or ‘unsupervised’). A number of tools from the Generic Model Organism Database consortium (GMOD, http://gmod.org/) can be used to predict gene models on assembled genome sequences. These tools usually identify repeats, align ESTs and proteins to a genome, produce *ab initio* gene predictions and automatically synthesize these data into gene annotations with evidence-based quality values. Although invaluable to the process of genome annotation, these tools have certain limitations and one may run the risk of erroneously perceiving them as mighty black boxes. Gene models can also be created

manually; you may have already done this for your work, running BLAST with your favorite gene.

Producing the official gene set (OGS) for a sequenced genome is often accomplished with one or more rounds of automatic gene building (de-novo or evidence-driven), the generation of a consensus set, and finally a community annotation effort (often through a distributed annotation system or via 'jamborees'). These community efforts on curation have very specific aims: to validate the automatic gene models, and to edit extant ones and/or create new ones as necessary. At the same time the community is asked to do a first pass on functional annotation by assigning terms to each model based on experimental data, the presence of protein domains, or sequence similarity to genes from model species.

As technical advances make sequencing faster and cheaper, the number of community annotation efforts has increased, augmenting a traditionally centralized model where all curators for a given genome project share the same physical location. This trend is particularly strong for smaller genome projects that rely largely on contributions from a geographically dispersed community of experts; this is the case of many of the projects that form i5K. WebApollo was designed to provide an easy to use, web-based environment that allows multiple distributed users to edit and share sequence annotations.

WebApollo (http://gmod.org/wiki/WebApollo) allows us to both edit gene models and assign functions. In genome sequencing projects, although the gene editing process is eventually stopped to produce a consensus set, the task of functional assignment continues and takes many years to be satisfactorily completed (as researchers perform experiments). Some species have dedicated annotation teams who scourge the literature for additional information (e.g. FlyBase at Cambridge UK, DictyBase at Chicago) but in the modern paradigm of NGS, there are too many new genomes and thus, a community-based effort might be more appropriate.

WebApollo is comprised of three components: i) a web-based client, ii) a server-side annotation editing engine, and iii) a server-side service that provides the client with data from different sources, including databases at the University of California at Santa Cruz, Ensembl, and Chado. The web-based client is designed as an extension to JBrowse, a JavaScript-based genome browser that provides a fast, highly interactive interface for the visualization of genomic data. The server-side annotation-editing engine is written in Java. It handles all the necessary logic for editing and deals with the complexities of modifications in a biological context, where a single change can have multiple cascading effects (e.g., when splitting or merging transcripts). The server provides synchronized updates over multiple browser instances, so that every edit is immediately visible to all users who are viewing or editing the same region. The server-side service that provides data to the client is built on top of Trellis, a Distributed Annotation System (DAS) server framework. All three components are open source and provided under the Berkeley Software Distribution (BSD) License. You may learn more about WebApollo on our Project web page at http://gmod.org/wiki/WebApollo, and by visiting with us at Poster EG-7 at the Arthropod Genomics Symposium.

During this WebApollo workshop you will join the WebApollo team for a hands-on demonstration of how to manually annotate automatically generated gene models. This Demo features the latest *Apis mellifera* genome assembly (Amel_4.5) and Official Gene Set

(amel_OGSv3.1). You will also have the opportunity to learn and practice manual annotation of your genes of interest while you test-drive the all New WebApollo. Thank you for joining us!

The WebApollo Team

Ed Lee[1], Gregg Helt[1], Nomi Harris[1], Robert Buels[2], Chris Childers[3], Justin Reese[3], Mónica Muñoz-Torres[3], Chris G Elsik[3], Ian Holmes[2] and Suzanna E. Lewis[1]

[1] Bioinformatics Open-source Projects, Lawrence Berkeley National Labs

[2] University of California, Berkeley

[3] Elsik Lab, Georgetown University

--

WebApollo documentation is available at http://gmod.org/wiki/WebApollo

Adapted from "Annotating a genome sequence: gene models". 2011 Next-gen sequencing Course, NESCent Academy. Alexie Papanicolaou and Monica Munoz-Torres. http://goo.gl/cDSR1.

What Annotators Should Look For

1. Annotating a simple case: **WHEN** "The official prediction is correct, or nearly correct, assuming that no aligned data extends beyond the OGS model and if so, it is not likely to be coding sequence, and/or the OGS prediction matches what you know about the gene":
    a. Can you add UTRs?
    b. Check exon structures.
    c. Check splice sites: …]5'-GT/AG-3'[…
    d. Check '*start*' and '*stop*' sites.
    e. Check the predicted protein product(s).
    f. If the protein product still does not look correct, go on to "*Annotating more complex cases*".

2. Additional functionality. You may also need to learn how to:
    a. Get genomic sequence
    b. Merge exons
    c. Add/Delete an exon
    d. Create an exon de novo (within an intron or outside existing annotations).
    e. Right/apple-click on a feature to get feature ID and additional information
    f. Looking up homolog descriptions going to the accession web page at UniProt/Swissprot

3. Annotating more complex cases:

   a. Incomplete annotation: protein integrity checks, indicate gaps, missing 5' sequences or missing 3' sequences.
   b. Merge of 2 OGS predictions on same scaffold
   c. Merge of 2 OGS predictions on different scaffolds **(uh-oh!).**
   d. Split of an OGS prediction
   e. Frameshifts, Selenocysteine, single-base errors, and other inconvenient phenomena

4. Adding **important** project information in the form of Canned and/or Customized Comments:
   a. NCBI ID, RefSeq ID, gene symbol(s), common name(s), synonyms, top BLAST hits (GenBank IDs), orthologs with species names, and **anything else you can think of**, because ***you*** are the expert.
   b. Type of annotation (e.g: whether or not the gene model was changed)
   c. Data source (for example if the Fgeneshpp predicted gene was the starting point for your annotation)
   d. The kinds of changes you made to the OGS gene model, e.g: split, merge
   e. Functional description
   f. Whether you would like for your MOD curator to check the annotation
   g. Whether part of your gene is on a different scaff

Always Keep in Mind

1. While at this workshop, and always when annotating gene models, remember a few things:
2. You are looking at a 'frozen photograph' of the genome assembly. You are not able to change that photograph; instead you will build a layer of information that may be placed on top so that it reflects the appropriate changes; much like photography-editing software would do to a picture.
3. The consensus set of automated predictions for the latest *Apis mellifera* genome assembly (Amel_4.5) was generated using GLEAN (Official Gene Set, amel_OGSv3.1). A gene model from the may be supported by any combination of models from the following sources: NCBI RefSeq, NCBI ab initio (GNOMON), Fgenesh, Fgenesh++, Augustus, NSCAN, and sgp2. You may also see evidence tracks from alignments with EST, gene models from other insect species, and transcriptomic data from a number of tissues and developmental stages; these alignments were calculated using algorithms like Exonerate or tBLASTx.
4. GLEAN and RefSeq entries constitute a high-quality gene prediction set; because of this, in many cases these models can be the starting point for your manual annotation efforts. When editing an existing gene model, please be sure to identify the corresponding RefSeq model (RefSeq ID: XP, NP for protein sequences; XM, NM for mRNAs) that you are planning to replace with your manual annotation.
5. In a number of cases you may choose not to edit a gene model; even if this is the case, it is of utmost importance that you check the corresponding RefSeq gene

model (if available). Even if no changes were made, you may also save your work and download the sequences.

6. There may be more than one transcript per GLEAN gene, so check them all. You may also wish to annotate additional transcripts rather than replacing a GLEAN transcript. In some cases, annotation may involve adding UTRs without modifying the CDS.
7. You will use WebApollo to verify or fix gene models, and to assign annotations. It is important to keep in mind that a documented approach, gathering supporting-evidence and quality control metrics, is the only way to ensure that the annotations will be of sufficient quality for a publication; remember to keep a log of everything you do. Keep also in mind that the process of automated annotation is not perfect: "incorrect and incomplete genome annotations will poison every experiment that uses them".
8. Manual curation uses the literature and public databases to infer gene function from experimental data and sequence-similarity searches within a phylogenetic framework. One should always drive manual curation efforts with a number of approaches: i) Sequence-similarity searches predict protein function. Distinguishing orthologs from paralogs helps to classify genes as members of the same family. Further, we must correct intron/exon and ORF boundaries, splice sites and frame-shift errors. It is often of interest to also identify promoters, polyadenylation sites and UTRs. Finally, structural (e.g. alternative splicing, inversions) and genetic variation (e.g. SNPs, indels) are all important for a number of downstream applications.

"**Annotate-Along" WebApollo Demo**

1. Log in at http://icebox.lbl.gov:8080/ApolloWebWorkshop-NUMBER/Login and replace with the sandbox number you were assigned. (1 through 50)

   e.g: http://icebox.lbl.gov:8080/ApolloWebWorkshop-51/Login

2. Username | Password: workshop

3. Find the region of interest:

   Group1.33 between coordinates 243,000 bp and 245,500 bp

4. Drag a few prediction tracks into the white-working area:

   Official Gene Set v3.1 (GLEAN)

   NCBI Gnomon

   SGP2

   Augustus Set 8

   NCBI RefSeq

   a. Find GLEAN model: GB40031-RA and RefSeq model: XM_003251648.1)
   b. Assess agreement among both predictions:
      - Do both models have the same number of exons?

- Do exon/intron boundaries match across tracks?

5. Drag evidence tracks onto the working area:
   NCBI ESTs
   Nurse RNA-Seq Coverage

   a. Inspect their coverage over predicted models, paying close attention to intron/exon boundaries of exon 3. These tracks of transcription evidence are useful to inform decisions on whether exon boundaries should be altered, and whether alternative splicing is evident, and isoforms should also be annotated.
   b. We will also take a few moments to drag and inspect the Nurse RNA-Seq Reads

6. Bring and modify models in the 'User-created Annotations' region.

   a. Drag the GLEAN and RefSeq models onto the 'User-created Annotations' track.
   b. Inspect the flagged exon ('!'): Could you correct the error?
   c. Drag evidence from the EST collection from NCBI, as well as the SGP2 track onto the "User-created" Annotations area. We will closely look at the region between 243900 and 244,250 to attempt to resolve the problems using both WebApollo tools, as well as BLAST alignments.

7. Feel free to browse around the Group1.33, or chose another group e.g: Group16.4. Try using the functions we just learned. We will be here to answer your questions and listen to your suggestions.