BMC Bioinformatics

POSTER PRESENTATION Open Access

Validation and quality assurance for genome browser database exports

Roger Chui[1], Jerzy W Jaromczyk[1*], Neil Moore[1], Christopher L Schardl[2]

From 14th Annual UT-KBRIN Bioinformatics Summit 2015
Buchanan, TN, USA. 20-22 March 2015

Background

A genome browser transition utility designed in our lab, FPD2GB2 (Fungal Project Database to GBrowse 2), exports data from a custom database used by the Fungal Endophytes Genome Project [1,2]. Designed as a collection of scripts, FPD2GB2 outputs the contents of a locally developed genome annotation database into the standard GFF3 format, allowing for bulk import of data into the GBrowse2 genome browser [3]. In short, FPD2GB2 is a collection of scripts designed to export data encoded in the Fungal Project Database format into a format which can be easily imported into GBrowse 2, namely GFF3.

Materials and methods

Any application which converts between data formats should ensure the completeness and accuracy of the output produced by FPD2GB2. Adding a data validator as part of the FPD2GB2 script collection allows for independent verification of the quality and soundness of the GFF3 files being imported into a production GBrowse2 environment.

We measure the accuracy of the output by comparing the features listed in the GFF3 files to the contents of the original database. Ensuring accurate offsets relative to reference features provides validation of accuracy. Comparing the parent-child inheritance structure of features in the output to that of the source data ensures the completeness of the output. The script collection is structured into a "master" script and several "worker" scripts, each of which produces its own output. The structure of the collection is shown in Figure 1. The goals and methods for the validator are described in Table 1.

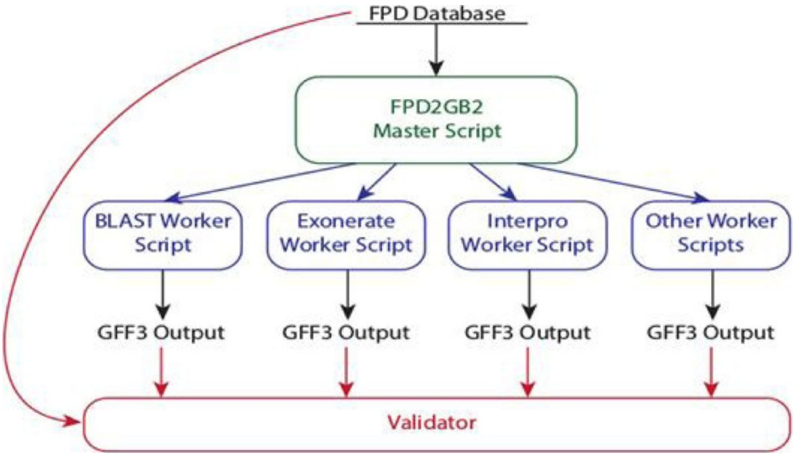

Figure 1 Block diagram of FPD2GB2 data flow and execution.

* Correspondence: jurek@cs.uky.edu
[1]Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA
Full list of author information is available at the end of the article

Table 1 Goals and methods for the validator.

***Completeness:* Verify that all data has been exported.**	Count base features in original database, ensure the count of features in output is an exact match.
***Correctness:* Ensure that the content exported accurately represents the original copy.**	Manual checking of a subset of features is a pragmatic way to ensure correctness, at least initially.
***Accuracy:* Check the placement of annotations relative to references to ensure there are no off-by-one errors.**	Select a set of features to use as reference. Calculate distance to other features in both FPD database and GFF3 file. Make sure that features in tracks that are directly related (e.g. exonerate results are derived from BLAST results) correspond at a high rate using a feature comparator such as ParsEval (part of the genometools package which is available at http://genometools.org/)
***Standards Compliance:* Verify output conforms to GFF3 standard http://www.sequenceontology.org/gff3.shtml**	Use gff3validator (also part of the genometools package) to ensure compliance with the GFF3 standard.

Results

It is notoriously difficult to prove accuracy of computational results and in practice validation is based on testing. In our case to validate the completeness, correctness and accuracy we use metrics which can not only give confidence that the output tends to accurately reflect the output, but also that the algorithms used to create the output are correct. The size of some of the databases and number of annotation tracks also makes full comparison of related tracks impractical, as fully comparing tracks takes a quadratic number of runs with respect to the number of tracks. Finally, because of the way the annotations do not have metadata establishing relationships, comparisons using ParsEval have to be run manually.

Authors' details
[1]Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA. [2]Department of Plant Pathology, University of Kentucky, Lexington, KY 40546, USA.

Published: 23 October 2015

References

1. Fungal Endophytes Genome Project :[http://www.endophyte.uky.edu/].
2. Schardl CL, Young CA, Hesse U, Amyotte SG, Andreeva K, Calie PJ, *et al*: **Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the Clavicipitaceae reveals dynamics of alkaloid loci.** *PLoS Genetics* 2013, **9(2)**:e1003323.
3. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, *et al*: **The Generic Genome Browser: A Building Block for a Model Organism System Database.** *Genome Research* 2002, **12(10)**:1599-1610.

doi:10.1186/1471-2105-16-S15-P13
Cite this article as: Chui *et al.*: **Validation and quality assurance for genome browser database exports.** *BMC Bioinformatics* 2015 **16**(Suppl 15): P13.