BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(8)

Tbl2KnownGene: A command-line program to convert NCBI.tbl to UCSC knownGene.txt data file

Yongsheng Bai

Department of Biology, Indiana State University, 600 Chestnut Street, Terre Haute, IN 47809, U.S.A; Yongsheng Bai - Email: Yongsheng.Bai@indstate.edu; *Corresponding author

Received August 07, 2014; Accepted August 08, 2014; Published August 30, 2014

Abstract:
The schema for UCSC Known Genes (knownGene.txt) has been widely adopted for use in both standard and custom downstream analysis tools/scripts. For many popular model organisms (e.g. Arabidopsis), sequence and annotation data tables (including "knownGene.txt") have not yet been made available to the public. Therefore, it is of interest to describe Tbl2KnownGene, a .tbl file parser that can process the contents of a NCBI .tbl file and produce a UCSC Known Genes annotation feature table. The algorithm is tested with chromosome datasets from Arabidopsis genome (TAIR10). The Tbl2KnownGene parser finds utility for data with other organisms having similar .tbl annotations.

Availability: Perl scripts and required input files are available on the web at http://thoth.indstate.edu/~ybai2/Tbl2KnownGene/index.html

Background:
To deposit or upload some gene records information into public popular databases requires users to conform the requirements of their standard feature tables. There is an increasing need in the bioinformatics field that standard tools should be able to convert tables annotated with certain formats to different structures.

The University of California Santa Cruz (UCSC) **[1]** Gene annotation files (e.g knownGene.txt) have been widely adopted by many standard and custom downstream analysis tools/scripts. The National Center for Biotechnology Information (NCBI) **[2]** .tbl file is a 5-column tab-delimited feature table containing genomic coordinate and other associated information of molecular records (gene, CDS, mRNA).

The Arabidopsis genome annotation files from TAIR **[3]** used .tbl table format to store gene annotation information. A universal tool for converting sequence and annotation data files from .tbl to "knownGene.txt" has not yet been developed. The author has developed a tool - Tbl2KnownGene, a .tbl file parser that can convert NCBI .tbl files and produce UCSC Known Genes annotation feature tables.

Methodology:
Download .tbl annotation files from the TAIR
The Arabidopsis .tbl file for each chromosome, a 5-column tab-delimited feature table containing genomic information of records (gene, CDS, mRNA), was downloaded from the TAIR (http://www.arabidopsis.org). A total of five chromosome files for Arabidopsis genome were obtained for the conversion. Each record contains the genomic coordinate start/end and other associated annotation information. See **Figure 1 and Table 1 (see supplementary material).**

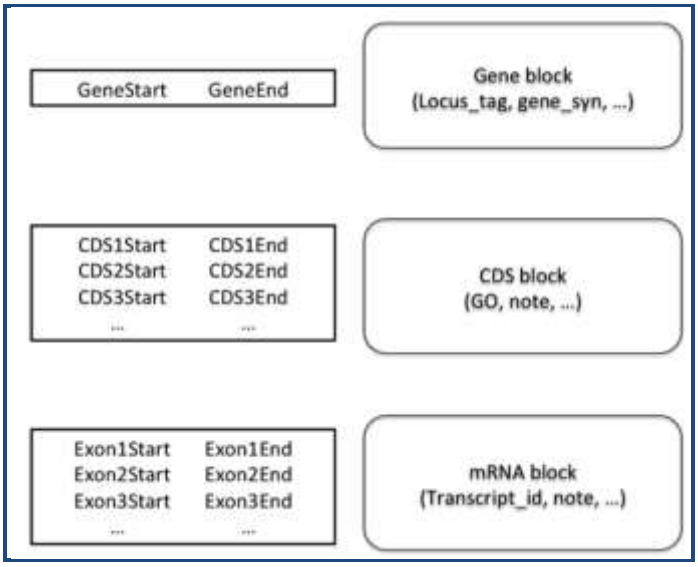

Figure 1: NCBI .tbl table annotation information

Convert .tbl to UCSC knownGene.txt:
The Tbl2KnownGene algorithm first classifies records into "blocks". Each block's contents are processed separately. The algorithm designates the leftmost start coordinate (rightmost start coordinate for "-") annotated for exons as the record start and the rightmost end coordinate (leftmost end coordinate for "-") as the record end. The algorithm concatenates all exon start locations for a transcript into a single comma-separated list, and likewise all exon end locations into a comma-separated list to comply with the UCSC knownGene schema format. The algorithm determines a gene's strand by comparing the record's start and end values. Since the UCSC knownGene.txt table always reports the exon coordinates in order from low to high, the algorithm reverses the order of the exon coordinates for genes coded on the negative strand. The pseudocode of the Tbl2KnownGene algorithm is shown in **Figure 2.** An example for a part of the input .tbl file of Arabidopsis is listed in **(Table 1)**.

Tbl2KnownGene Input and Output:
The input are .tbl files (e.g. the chromosome files of Arabidopsis) and the output are annotated UCSC KnownGene.txt files. A truncated example file is shown in **Table 2 (see supplementary material)**.

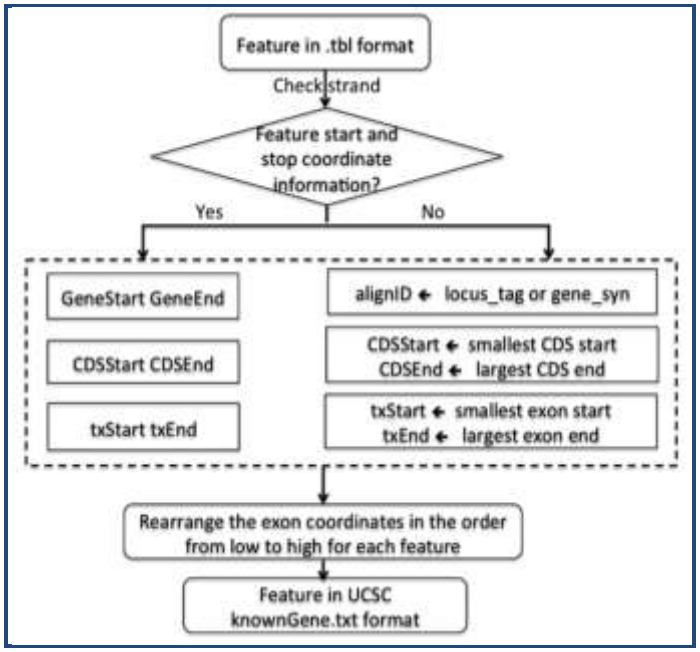

Figure 2: The pseudocode of Tbl2KnownGene algorithm

Conclusions:
Efficient pipelines/tools are needed for downstream analyses for next-generation sequencing data. Even though UCSC database tables have been built for many organisms/species, the research community requires similar annotations for other organisms. A PERL parser named Tbl2KnownGene converts the contents of a NCBI .tbl annotation table to a UCSC Known Genes annotation table used by other downstream analysis.

Acknowledgement:
The author thanks reviewers for comments and Dr. James Cavalcoli at University of Michigan for suggestions.

References:
[1] http://genome.ucsc.edu
[2] http://www.ncbi.nlm.nih.gov/genbank
[3] http://www.arabidopsis.org/

Edited by P Kangueane
Citation: **Bai,** Bioinformation 10(8): 544-547 (2014)

Supplementary material:

Table 1: A part of the input file in NCBI .tbl format

```
>Feature ref|NC_003070|
 3631    5899    gene
                 locus_tag      AT1G01010
                 db_xref TAIR:AT1G01010
                 gene_syn       ANAC001, NAC domain containing protein 1, NAC001, T25K16.1, T25K16_1
                 gene    NAC001
 3760    3913    CDS
 3996    4276
 4486    4605
 4706    5095
 5174    5326
 5439    5630
                 go_process     regulation of transcription|GO:0045449||IEA
                 go_component    cellular_component|GO:0005575||ND
                 go_process     multicellular organismal development|GO:0007275||ISS
                 go_function    sequence-specific DNA binding transcription factor activity|GO:0003700|11118137|ISS
                 product NAC domain containing protein 1
                 note    NAC domain containing protein 1 (NAC001); FUNCTIONS IN: sequence-specific DNA binding transcription
factor activity; INVOLVED IN: multicellular organismal development, regulation of transcription;
LOCATED IN: cellular_component unknown; EXPRESSED IN: 7 plant structures; EXPRESSED DURING: 4 anthesis, C globular
stage, petal differentiation and expansion stage; CONTAINS InterPro DOMAIN/s: No apical meristem (NAM) protein
(InterPro:IPR003441); BEST Arabidopsis thaliana protein match is: NAC domain containing protein 69 (TAIR:AT4G015
 50.1); Has 2503 Blast hits to 2496 proteins in 69 species: Archae - 0; Bacteria - 0; Metazoa - 0; Fungi - 0; Plants - 2502; Viruses - 0;
Other Eukaryotes - 1 (source: NCBI BLink).
                 protein_id     AT1G01010.1p
                 transcript_id  AT1G01010.1
                 locus_tag      AT1G01010
 3631    3913    mRNA
 3996    4276
 4486    4605
 4706    5095
 5174    5326
 5439    5899
                 protein_id     AT1G01010.1p
                 transcript_id  AT1G01010.1
                 locus_tag      AT1G01010
                 note    supported by cDNAs: gi|24796987|gb|BT001115|, gi|16612276|gb|AF439834|, gi|110742029|gb|AK226863|;
supported by ESTs: gi|49272440|gb|BP621258|, gi|58800648|g
 b|BP779869|, gi|19868695|gb|AV826635|, gi|125053743|gb|EL124732|, gi|19844129|gb|AV810144|, gi|59266962|gb|BP788378|,
gi|56087255|gb|BP562468|, gi|19829901|gb|AV795918|
```

Table 2: A part of the output file in KnownGene.txt format

#name chrom strand txStart txEnd cdsStart cdsEnd exonCount exonStarts exonEnds proteinID alignID
AT1G01010.1 chr1 + 3631 5899 3760 5630 6 3631,3996,4486,4706,5174,5439, 3913,4276,4605,5095,5326,5899, AT1G01010.1p ANAC001, NAC domain containing protein 1, NAC001, T25K16.1, T25K16_1
AT1G01020.1 chr1 - 5928 8737 6915 8666 10 5928,6437,7157,7384,7564,7762,7942,8236,8417,8571, 6263,7069,7232,7450,7649,7835,7987,8325,8464,8737, AT1G0
1020.1p ARV1, T25K16.2, T25K16_2
AT1G01020.2 chr1 - 6790 8737 7315 8666 8 6790,7157,7564,7762,7942,8236,8417,8571, 7069,7450,7649,7835,7987,8325,8464,8737, AT1G01020.2p ARV1,
T25K16.2, T25K16_2
AT1G01030.1 chr1 - 11649 13714 11864 12940 2 11649,13335, 13173,13714, AT1G01030.1p NGA3, NGATHA3, T25K16.3, T25K16_3
AT1G01040.1 chr1 + 23146 31227 23519 31079 20
23146,24542,24752,25041,25524,25825,26081,26292,26543,26862,27099,27372,27618,27803,28708,28890,29160,30147,30410,309
02,
24451,24655,24962,25435,25743,25997,26203,26452,26776,27012,27281,27533,27713,28431,28805,29080,30065,30311,30816,312
27, AT1G01040.1p ABNORMAL SUSPENSOR 1, ASU1, ATDCL1, C
AF, CARPEL FACTORY, DCL1, DICER-LIKE 1, EMB60, EMB76, EMBRYO DEFECTIVE 60, EMBRYO DEFECTIVE 76, SHORT INTEGUMENTS 1, SIN1, SUS1, SUSPENSOR 1, T25K16.4, T25K16_4, dicer-like 1
AT1G01040.2 chr1 + 23416 31120 23519 31079 20
23416,24542,24752,25041,25524,25825,26081,26292,26543,26862,27099,27372,27618,27803,28708,28890,29160,30147,30410,309
02,
24451,24655,24962,25435,25743,25997,26203,26452,26776,27012,27281,27536,27713,28431,28805,29080,30065,30311,30816,311
20, AT1G01040.2p ABNORMAL SUSPENSOR 1, ASU1, ATDCL1, C
AF, CARPEL FACTORY, DCL1, DICER-LIKE 1, EMB60, EMB76, EMBRYO DEFECTIVE 60, EMBRYO DEFECTIVE 76, SHORT INTEGUMENTS 1, SIN1, SUS1, SUSPENSOR 1, T25K16.4, T25K16_4, dicer-like 1
AT1G01050.1 chr1 - 31170 33153 31382 32670 9 31170,31521,31693,31933,32088,32282,32431,32547,33029, 31424,31602,31813,31998,32195,32347,32459,32670,33153, AT1G0
1050.1p AtPPa1, PPa1, T25K16.5, T25K16_5, pyrophosphorylase 1
AT1G01060.1 chr1 - 33666 37840 33992 37061 9 33666,34401,35567,35730,36624,36810,37023,37373,37569, 34327,35474,35647,35963,36685,36921,37203,37398,37840, AT1G0
1060.1p LATE ELONGATED HYPOCOTYL, LATE ELONGATED HYPOCOTYL 1, LHY, LHY1, T25K16.6, T25K16_6