

GBKFinisher: A tool for GenBank files refinement

Gustavo Santos de Oliveira¹, Douglas Parise¹, Mariana Teixeira Dornelles Parise¹,
Anne Cybelle Pinto Gomide², Vasco Ariston de Carvalho Azevedo¹,

1 Universidade Federal de Minas Gerais

2 Federal University of Minas Gerais

Abstract

After genome assembly and annotation, usually a lot of effort is taken to check for potential pseudogenes as well as to set annotation features ready to submission. Those processes are usually manually performed in Artemis and are time consuming. Different filters set in Artemis help in the identification of potential issues, as absence of stop codons in CDSs, incorrect start codons, duplication and overlap of features, etc. For potential pseudogenes, however, a manual inspection must be performed. The user must check for frameshifts corrections by observing for possible indels, the presence of premature stop codons or even gene fragments. The user, then frequently uses BLASTp as a guide for possible correction of those different issues. Based on the need for a more automatic process we introduce GBKFinisher. This tool was developed as an effort to help users to save time and more accurately fix annotation issues. GBKFinisher is based on three major packages: GBKChecker, GBKParser and GBKSolver. GBKChecker attempts to give the basic stats, like qualifiers count and nucleotide composition, as well as a diagnosis of possible issues, like the number of possible frameshifts, fragments and split ORFs that could be safely joined, without a stop codon in between. GBKParser was created to rewrite gbk files with qualifiers as specified by the user. GBKSolver attempts to automatically solve for detected issues by GBKChecker. GBKFinisher allows a very convenient way for users to filter out unwanted qualifiers, as well as to annotate locus_tag differently for CDSs, tRNAs, and rRNAs. It is also possible for users to manually correct potential issues detected in GBKChecker. A log file with all possible issues encountered by GBKChecker is maintained and can be accessed for detailed troubleshooting. Likewise, a gbk file is created with a color scheme that assists users in the curation process. The primary output of GBKFinisher is a user defined Genbank file, that can ultimately be used for guiding the curation process or even for sequin and GenBank submission.

Funding: Fapemig