Research Article

Open Access

[1] Department of Genome Mapping, Molecular Genetics and Genome Mapping Laboratory, Agricultural Genetic Engineering Research Institute, Giza, Egypt.
[2] International Center for Agricultural Research in the Dry Areas (ICARDA), Cairo, Egypt.

Contacts of Authors

* To whom correspondence should be addressed: Alsamman M. Alsamman.

GeneSyno : Simple tool to extract gene sequence from the human genome despite synonymous gene terms

Alsamman M. Alsamman[1]* , Peter T. Habib[2]

Abstract

Extracting gene data from the human genome is a tricky task. Gene name is the key information for harvesting its sequence, annotation, and other related data.Unfortunately, most human genes have different and multiple names, depending on the database and the resource in which they have been published. Such an issue is delaying the ability of researchers to gather the necessary knowledge and to build their opinion on the function of genes. Here we introduce GeneSyno, a simple, versatile and reliable tool that can be used to extract gene information from human genome data even though it is synonymous gene names. GeneSyno was written using C and Python programming languages and could easily be integrated into another pipeline.

Keywords: gene information, human genome, gene name , gene annotation, gene, synonymous gene name.

Background

Human Genome Research is one of the most intensive research fields. Synonymous terms of the name of the gene remain a major issue in human genomics (1). Several human genes have different names, depending on the databases, the articles or the newly discovered function. With even more research articles published online, this information has become challenging for efficient implementation and reuse (2). Such a case has been a complicated issue where genomic scientists can not form a collective and prospective conclusion by using published information on most human genes.

Several tools have been published to solve this problem, where text mining, and searching for databases could be used to generate symbol co-occurrences to extend information extraction capabilities (3-6).The main problem that, most of these tools require high computational skills, or are available only in online versions. Most of these tools require high computational skills, or are only available in online versions. This may constrain the ability of researchers to access all available information for massive lists of genes at any time.

Here we introduce GeneSyno, a simple, versatile and reliable tool that can be used to extract gene information from human genome data even though it is synonymous gene names. GeneSyno was written using C and Python programming languages and could easily be integrated into another pipeline.

Material and methods

GeneSyno was built using C and Python3 programming languages. The user's input will be a list of human gene names. The input from the user will be a list of names for human genes. GeneSyno collects all available information about these genes from the GRCh38 database (which users could change for newer versions) and reports a tab-limited file containing gene information such as gene name, official gene name, chromosome , description, gene start, gene end, and a list of gene synonym names. Furthermore, it produced a FASTA file that contains sequences of all genes' proteins . If gene have more than one protein (isoforms) it will reported (**Figure 1**).

GeneSyno core was written using C programming language to extract massive lists of information about genes in less processing time. GeneSyno can be installed and used on various operating systems, and has a simple GUI for users.

Figure 1 : Example of GeneSyn input and outputs. If a list of names of human genes is given in text format (A), GeneSyno can produce a table containing all gene information in tab delimited format (B) and protein sequences in FASTA formats (C) .

Availability

GeneSyno is available as a standalone tool at : https://github.com/AlsammanAlsamman/GeneSyno

References

1. Cohen KB, Acquaah-Mensah GK, Dolbey AE, Hunter L. Contrast and variability in gene names. In: Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3. 2002. p. 14–20.

2. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 2005;21(2):248–56.

3. Cohen AM, Hersh WR, Dubay C, Spackman K. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinformatics. 2005;6(1):103.

4. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12(1):32–42.

5. Girish K, Dubey S. Eukaryotic Molecular Biology Databases: An Overview. Highlights Biosci. 2018;1:1–7.

6. Alsamman AM. The Art of Bioinformatics Learning in Our Arabic World. Highlights Biosci. 2019;2.