BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software
Volume 8(25)

Sequence Maneuverer: tool for sequence extraction from genomes

Tayyaba Yasmin[1]*, Inayat Ur Rehman[2], Adnan Ahmad Ansari[1], Khurrum liaqat[1] & Muhammad Irfan khan[1]

[1]Department of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan; [2]Department of Computer Sciences, COMSATS Institute of Information Technology, Islamabad, Pakistan; Tayyaba Yasmin - Email: drtayyabayasmin@gmail.com; *Corresponding author

Received March 21, 2012; Revised October 16, 2012; Accepted December 04, 2012; Published December 19, 2012

Abstract:

The availability of genomic sequences of many organisms has opened new challenges in many aspects particularly in terms of genome analysis. Sequence extraction is a vital step and many tools have been developed to solve this issue. These tools are available publically but have limitations with reference to the sequence extraction, length of the sequence to be extracted, organism specificity and lack of user friendly interface. We have developed a java based software package having three modules which can be used independently or sequentially. The tool efficiently extracts sequences from large datasets with few simple steps. It can efficiently extract multiple sequences of any desired length from a genome of any organism. The results are crosschecked by published data.

Availability: URL 1: http://ww3.comsats.edu.pk/bio/ResearchProjects.aspx
URL 2: http://ww3.comsats.edu.pk/bio/SequenceManeuverer.aspx

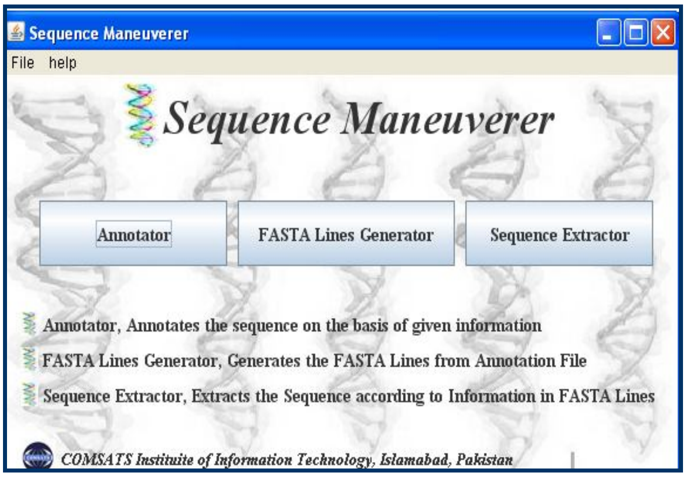
Key Words: Annotation, Biology and Genetics, Bioinformatics Software, Coding tools and Techniques

Background:

The DNA sequences of many organisms are available through different databases. Moreover, the easy access of many sequenced genomes has enhanced the pace of research in the field of bioinformatics **[1].** Analysis of coding **[2]** and non-coding **[3]** regions of some of the genomes has revealed the underlying biological messages to some extent but this is like touching the tip of the iceberg. Biological interpretations of non coding sequences are rather more challenging due to their abundance and nonspecific pattern of occurrence in genomes **[4].** Genome wide studies usually require extraction of large DNA sequences from a given data set. Normally, DNA sequences are stored in specific formats in different databases and users extract the related information according to the experimental objectives. Different softwares available for DNA sequence extractions have their own pros and cons and no single software can fulfill all the requirements of a user at one time **[5-8].** The extraction of coding/non coding sequences from the chromosome files stored in a database is a vital and basic step in research plans in the field of bioinformatics. Sequence maneuverer has been designed to resolve this problem which takes GenBank file as input and generates FASTA lines. These FASTA lines are used as input in sequence extractor which then extracts the sequences accordingly.

Methodology:

Sequence maneuverer basically consists of three modules named as annotator, FASTA line generator and sequence extractor. These modules could be used independently or in combination depending upon the user `s objectives. The main interface is shown in (**Figure 1).** This software has been implemented in Java programming language. A system requirement for this software is Java Virtual Machine (JVM). The annotator deals with annotation files available in GenBank formats. FASTA Lines Generator creates FASTA lines and writes them in a text file which can be used as the input file for sequence extractor. The software will extract the sequence efficiently through sequence extractor.



Annotator, Annotates the sequence on the basis of given information

FASTA Lines Generator, Generates the FASTA Lines from Annotation File

Sequence Extractor, Extracts the Sequence according to Information in FASTA Lines

COMSATS Instituite of Information Technology, Islamabad, Pakistan

Figure 1: The interface of Sequence Maneuverer.

Fasta Lines Generator

The user can specify different attributes like, project name, authors name and the project details. The resulting information will be stored in a separate file named as Project Details.txt. "Browse" button takes input from GenBank formatted files and then by clicking on the "generate" button the user can get FASTA file named as "FASTAz.txt". This text file (FASTAz.txt) contains FASTA lines for the annotation file of any chromosome the user has specified as input.

Sequence Extractor

The software package deals with the FASTA lines and the chromosome files that user chooses in order to extract the dataset for sequence analysis. Currently, there is a shortage of publically available stand alone applications for extraction of sequence upstream or downstream of the transcription start site (TSS) or coding DNA sequences (CDS) that uses the FASTA lines. An effort has been made in this regard; a desktop application has been developed with a user friendly interface. Moreover, its efficiency and effectiveness is evident from its fast extraction process without RAM-intensive file loading operations. The **Table 1 (see supplementary material)** shows the system specification, extraction time and other details. Work Flow of Sequence Extractor is shown in (**Figure 2).**

Validation:

The output generated by this program was tested manually by checking the TSS location and its sequences from Arabidopsis thaliana genome. Extracted sequence length was about 200 upstream and 50 downstream. The output was set of 251 nucleotides long sequences (TSS at +1). Furthermore, comparison of the output with publically available datasets revealed that our results substantially matched with the output of published datasets (http://linux1.softberry.com/data/plantprom/Links/PLPR_predicted_ATceres.seq). In addition to the validation of promoter sequences, CDS results of this software also matched with the results obtained from NCBI.

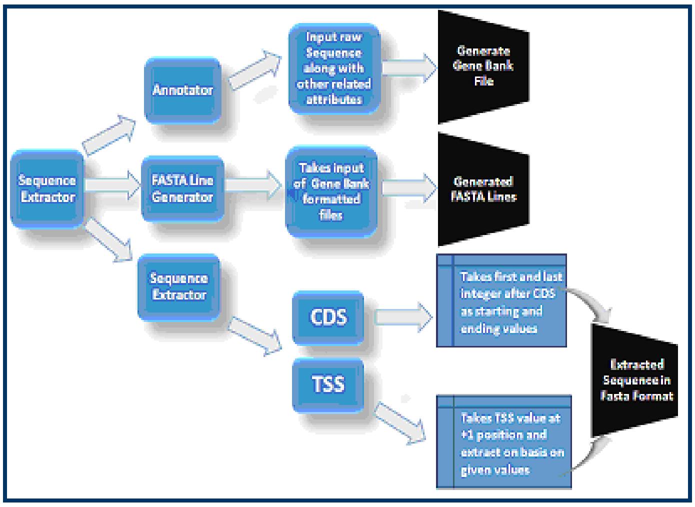

Sequence Extractor
Annotator
FASTA Line Generator
Sequence Extractor
Input raw Sequence along with other related attributes
Takes input of Gene Bank formatted files
CDS
TSS
Generate Gene Bank File
Generated FASTA Lines
Takes first and last integer after CDS as starting and ending values
Takes TSS value at +1 position and extract on basis on given values
Extracted Sequence in Fasta Format

Figure 2: Work flow of Sequence extractor**.**

Utility:

Efficient sequence extraction of any desired length from genome of any organism; multiple sequences can be handled or manipulated simultaneously; any raw sequence can be converted into GenBank format using annotator.

Acknowledgement:

The guidance provided by Dr. Ilham Shahmuradove (Head of Bioinformatics Lab, Institute of Botany, Azerbaijan National Academy of Sciences, Baku, Azerbaijan), Mr. Nadeem Malik (Department of Biosciences, CIIT, Islamabad, Pakistan) and Aahid Naeem Supra (Department of Biosciences, CIIT, Islamabad, Pakistan) is highly acknowledged.

References:

[1] White WTJ & Hendy MD, *BMC Bioinformatics*. 2008 **9**: 242 [PMID: 18489794]
[2] Yin C & Yau SS, *J Theor Biol*. 2007 **247**: 687 [PMID: 17509616]
[3] Alexander RP *et al. Nat Rev Genet*. 2010 *11*: 559 [PMID: 20628352]
[4] Buchanan CD *et al. Genetics*. 2004 **168**: 1639 [PMID: 15579713]
[5] Sun Q *et al. Nucleic Acids Res*. 2009 **37**: D969 [PMID: 18832363]
[6] Thomas-Chollier M *et al. Nucleic Acids Res*. 2008 36: W119 [PMID: 18495751]
[7] http://whitman.myweb.uga.edu/detools.html
[8] Halees A *et al. Nucleic Acids Res*. 2003 **31**: 3554 [PMID: 12824364]

Edited by P Kangueane
Citation: Yasmin *et al.* Bioinformation 8(25): 1277-1279 (2012)

Supplementary material:

Table 1: Salient features of Sequence Extractor

System Specification	Number of FASTA Lines to be extracted	Extracted Sequence Length	DNA Strand	Reverse Complement	Average Time*
1 GB RAM, 2.33 GHz Dual core	241	2001	Positive	No	1 min 59 sec
1 GB RAM, 2.33 GHzDual core	241	2001	Negative	Yes	2 min 0 sec

* Time is calculated using Net beans IDE timer.