APPLICATION NOTE Genomics & Informatics Vol. 7(4) 217-219, December 2009

ORF Miner: a Web-based ORF Search Tool

Sin-Gi Park and Ki-Bong Kim*

Department of Medical Biotechnology, Sangmyung University, Cheonan 330-720, Korea

Abstract

The primary clue for locating protein-coding regions is the open reading frame and the determination of ORFs (Open Reading Frames) is the first step toward the gene prediction, especially for prokaryotes. In this respect, we have developed a web-based ORF search tool called ORF Miner. The ORF Miner is a graphical analysis utility which determines all possible open reading frames of a selectable minimum size in an input sequence. This tool identifies all open reading frames using alternative genetic codes as well as the standard one and reports a list of ORFs with corresponding deduced amino acid sequences. The ORF Miner can be employed for sequence annotation and give a crucial clue to determination of actual protein-coding regions.

Availability: ORF Miner is available at http://203.230.194.163/sgpark/orfminer.html

Keywords: alternative genetic codes, gene prediction, open reading frame, ORF Miner, protein-coding region

Introduction

As a general thing, the primary challenge that follows the sequencing of anything from a small segment of DNA to a complete genome is to assign as much information as possible to the raw sequence data, i.e., to establish where the various of functional elements such as genes, promoters, terminators etc., lie in the sequence. Such a task called sequence annotation is an integral part of functional genomics. It is a rapidly evolving field in genomics, driven predominantly by computational tools (Brent, 2008). The annotation task can be classified into two steps: Structural annotation, which deals with identification of biologically relevant sites in the sequence, and functional annotation, which attributes specific biological information to the sites found in the first step. The most crucial step in structural annotation is to determine protein-coding regions. Even though there are many gene finding programs with sophisticated and intricate algorithms (Burge, 1997, Delcher, 2007), relatively straightforward and powerful way to find out protein-coding regions is to search for open reading frames (ORFs).

The open reading frame is a portion of an organism's genome containing a sequence of bases that could potentially encode a protein. In other word, it means that the region certainly looks like a gene, but it has not been proved to actually be a gene. In a gene, ORFs are located between the start code sequence (initiation codon) and stop code sequence (termination codon) so that the existence of an ORF, especially a long one, can be usually a good indication of the presence of a gene in the surrounding sequence. However, ORFs are usually encountered when sifting through pieces of DNA while trying to locate a gene. In addition, ORFs can also occur by chance outside of genes. The situation in prokaryotes is relatively straightforward since scarcely any eubacterial and archaeal genes contain introns. The situation is much more complicated in eukaryotes where the majority of genes are composed of introns and exons, and further analysis must be required to detect the intron/exon boundaries and assemble the exons into a contiguous coding sequences. The Coding Sequence (CDS) and the ORF may be interchangeable but they are a little different from each other. The CDS is the actual region of DNA that is translated to form proteins. In Eukaryotes, while the ORF may contain introns as well, the CDS refers to contiguous coding sequence or concatenated exons that can be divided into codons. In Prokaryotes, the ORF and the CDS can be considered the same entity.

As mentioned above, the ORF can be a good indication of the presence of a CDS or a gene and the determination of ORFs is the first step toward the gene prediction, especially for prokaryotes. Furthermore, ample ORFs of specific organisms are required to build a general probabilistic model of the gene structural and compositional properties of genomic DNA sequences of the corresponding organisms, which is introduced and applied to the problem of identifying genes in unannotated genomic sequences. In terms of sequence annotation, it may be reasonable that all possible ORFs should be provided regardless of high specificity and the curator should decide on which ORFs are actually coding regions. In these respects, we have developed a web-based ORF search tool, the ORF Miner, which fo-

*Corresponding author: E-mail kbkim@smu.ac.kr
Tel +82-41-550-5377, Fax +82-41-550-5184
Accepted 30 November 2009

cuses on identifying literal ORF based on genetic codes.

Overview and Features of ORF Miner

The ORF Miner is implemented in PHP and GD graphics library on a PC server of which the operating system is Red Hat Linux 3.4.3-9.EL4. It can run on any Linux or Unix system that has PHP version 5.2 or higher and GD version 2.0 or higher alike installed and be available via a Apache web server. The client requests the web server to run a PHP-CGI server, the ORF Miner, with a set of input parameters. In response to the request, the web server runs the ORF Miner and receives the output results that will be transmitted to be finally displayed in the client (Fig. 1).

The parameters for this tool consist of input sequence, genetic code, minimum ORF size, and display data type (Fig. 2). The input sequence can be directly cut and pasted into the edit box or be uploaded from a file. Both of FASTA format and plain text format are allowed for the input sequence. In case of genetic code parameter, 17 genetic codes are provided to be exclusively chosen among them. This tool identifies all open reading frames using the standard or alternative genetic codes. Moreover, the parameter "Output sequence type" can be set to one of 'Nucleotide', 'Peptide' or 'Both' and the user can specify the minimum ORF size as a parameter. A minimum ORF size prevents short ORFs from being detected and displayed.

The strategy of ORF Miner to detect ORFs can be summarized as follows:

a. Check if all the parameters are set correctly before proceeding forward. An error message will be displayed in case there is something wrong with any parameter setting.

b. Read the input DNA sequence in six reading frames. The input sequence will be considered as the forward strand. Three are in the forward and three in the reverse direction.

c. Use foreach loop to scan the DNA sequence for start and stop codons in each of the six reading frames on both strands using pattern matching (regular expressions). Start and stops codons are defined by the genetic code specified by the user.

d. Store the coordinate, nucleotide sequence and translated protein sequence of the detected ORF of length higher than or equal to "minimum OFR length" value into a hashing table (an associative array) for corresponding reading frame.

e. Draw a figure with GD library, which displays the relative positions of all detected ORFs and report a list of ORFs.

The analysis output looks like the one in Fig. 3. The upper part of the analysis output contains the figure drawn up with GD library. It has both shaded and unshaded portions. The shaded portions correspond to ORFs. A list of all detected ORFs is displayed as a FASTA format under the output figure. In the analysis output, (1), (2), and (3) refer to the first, the second, and the third reading frame of the forward strand respectively and (-1), (-2), and (-3) refer to corresponding reading frame of the reverse strand. The symbols (+) and (-) indicate forward strand and the reverse strand respectively. The ORFs less than "minimum ORF length" value will not be shown in the analysis output. The comment line in a FASTA format in a list of ORFs provides useful information on each ORF, including reading frame, strand, coordinate (start, end), and length.

Discussion

The ORF Miner is a graphical analysis utility which de-

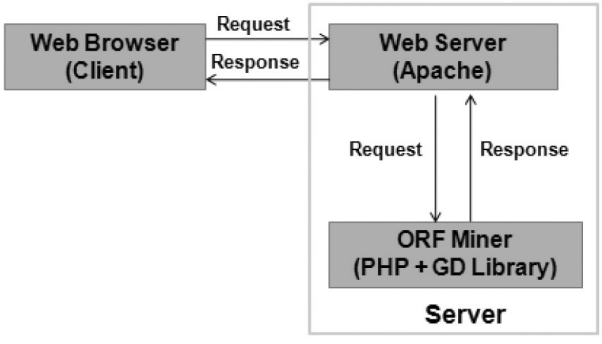

Fig. 1. Overall architecture of the ORF Miner.

Fig. 2. Web interface of the ORF Miner.

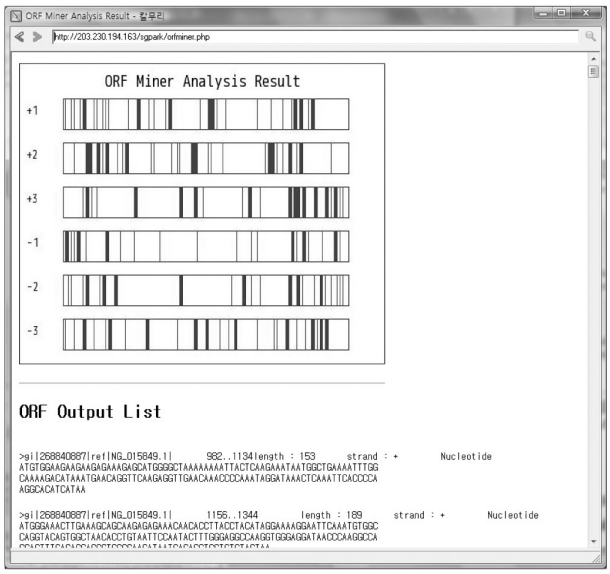

Fig. 3. Analysis output of the ORF Miner.

termines all possible open reading frames of a selectable minimum size in an input sequence. This tool identifies all open reading frames using alternative genetic codes as well as the standard one and reports a list of ORFs with corresponding deduced amino acid sequences. Having a list of ORFs provided by this tool, the end-user or the curator can contribute to rejecting unlikely ORFs and including likely ORFs for the purpose of sequence annotation and further experimental design. As mentioned in the introduction, this work concentrates on identifying literal ORFs based on genetic codes, not detecting which ORFs actually encode proteins. The latter is another work. There are a number of difficulties in determining if an ORF is actually used by the organism to code for protein. A powerful, but not often available, indication is conservation across related species. In addition, some of these problems can be resolved if all the available sequence information is combined, including sequence homology, possible promoters, initiation, translation signals and transcription terminators. In this context, through further work, we will incorporate some of sequence information to facilitate users to identify which of the six frames actually encodes the protein or actual ORF among a list of ORFs.

Acknowledgements

This work was supported financially by RIET (Research Institute of Engineering Technology) of Sangmyung University.

References

Brent, M.R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews Genetics* 9, 62-73.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78-94.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23, 673-679.