Research Journal of Recent Sciences _____ ISSN 2277-2502
Vol. **1(11),** 32-35, November **(2012)** *Res.J.Recent Sci.*

ORF Investigator: A New ORF finding tool combining Pairwise Global Gene Alignment

Dwivedi Vivek Dhar[1]* and Mishra Sarad Kumar[2],
[1]Department of Bioinformatics, UCST, Dehradun, INDIA
[2]Department of Biotechnology, DDU University, Gorakhpur, INDIA

Available online at: www.isca.in
Received 10th September 2012, revised 14th September 2012, accepted 21st September 2012

Abstract

Bioinformatics tools have become an integral part of the molecular data generated during the DNA fingerprinting of fungal pathogens. Finding and annoting the coding and non coding regions and final product in the form of its amino acid sequences is prerequisite for understanding the evolutionary processes in different pathogenic, fungi, as well as the species used for bioremediation, the medicinal and for biofertilizers applications. In the present study an attempt has been made to develop a tool "ORF Investigator" which not only gives information about the coding and non coding sequences but also can perform pairwise global alignment of different gene/DNA regions sequences. The tool efficiently finds out the ORFs for corresponding amino acid sequences and converts them into their one letter amino acid code declaring their respecting positions in the sequence stretch. The pairwise global alignment between the sequences makes it convenient to detect the different mutations including single nucleotide polymorphism. Needleman and Wunsch algorithms are used for the gene alignment and the coding has been done in PERL language making it suitable for windows user.

Keywords: ORF, alignment, perl, investigator, DNA.

Introduction

Research in the biosciences increasingly depends upon bioinformatics for the effective analysis of biological data and experimental results[1]. Having access to the appropriate bioinformatics tools is crucial to the success of any research project. The field of bioinformatics itself has previously been segregated into the parallel realms of data analysis[2-3]. As a result, a large proportion of bioinformatics work involves the analysis of data from a variety of sources through a pipeline of analysis tools. Identifying the coding regions with their amino acid translations in the DNA sequences and alignment of two genes/DNA regions sequences to find the homology between them opens door for bioinformatics research[4]. The simplest method of finding DNA sequences that encode proteins is to search for open reading frames, or ORFs. An ORF is a length of DNA sequence that contains a contiguous set of codons, each of which specifies an amino acid. There are six possible reading frame sin every sequence, three starting at positions 1, 2, and 3 and going in the 5´ to 3´ direction of a given sequence, and another three starting at positions 1, 2, and 3 and going in the 5´ to 3´ direction of the complementary sequence[6]. In prokaryotic genomes, DNA sequences that encode proteins are transcribed into mRNA, and the mRNA is usually translated directly into proteins without significant modification[7]. The longest ORFs running from the first available Met codon on the mRNA to the next stop codon in the same reading frame generally provide a good, but not assured prediction of the protein-encoding regions. A reading frame of a genomic sequence that does not encode a protein will have short ORFs due to the presence of many in-frame stop codons. In eukaryotic organisms, transcription of protein-encoding regions initiated at specific promoter sequences is followed by removal of noncoding sequence (introns) from pre mRNA by a splicing mechanism, leaving the protein-encoding exons. Once the introns have been removed and certain other modifications to the mature RNA have been made, the resulting mature mRNA can be translated in the 5´ to 3´ direction, usually from the first start codon to the first stop codon[9]. As a result of the presence of intron sequences in the genomic DNA sequences of eukaryotes, the ORF corresponding to an encoded gene will be interrupted by the presence of introns that usually generate stop codons. ORF Investigator involves in finding the open reading frames (ORFs) and performing Pairwise global alignment between two genes and DNA regions sequences. The global alignment is stretched over the entire sequence length to include as many matching nucleotides as possible up to and including the sequence ends. ORF Investigator was created to provide a general framework for research-focused bioinformatics tasks to overcome these challenges and take advantage of modern computing trends.

Material and Methods

ORF Investigator is written in Perl language to maximize interoperability among all commonly used operating systems. Perl Tk is used for windows programming for better looks. It is compiled under the Perl interpreter and converted into executable file using Perl2exe program. This Program uses the simple biological algorithms for ORF detection[10]. First it takes DNA sequence as input and translates them into six possible

reading frames and detects the ORFs, ORFs length, ORFs positions and its final product in the form of amino acid sequences with their length. For Pairwise global gene alignment it uses the dynamic programming algorithms of optimal alignment between two DNA sequence given by Needleman and Wunsch. This method compares every pair of characters in the two DNA sequences and generates an alignment. This alignment will include matched and mismatched characters and gaps in the two sequences that are positioned so that the number of matches between identical or related characters is the maximum possible. The dynamic programming algorithm provides a reliable computational method for aligning DNA sequences. The method has been proven mathematically to produce the best or optimal alignment between two sequences under a given set of match conditions[11]. Optimal alignments provide useful information to biologists concerning sequence relationships by giving the best possible information as to which characters in a sequence should be in the same column in an alignment, and which are insertions in one of the sequences (or deletions on the other)[12]. This information is important for making functional and evolutionary predictions on the basis of sequence alignments.

Results and Discussion

The ORF Investigator[13] is a graphical user interface program comprises a sequence text area for sequence, four pull-down menus and many windows right-click functions for common bioinformatics analyses. First menu named file has six options – open, new, save, save as, clear all, and quit respectively. Using open option a DNA fasta format sequence file of interest from any directories of your hard disk can be open. New option remove previous file and provide space for new fasta sequence file entry. Save option save your modified fasta file with same file name. Save as option gives a chance to save your file with new name. Clear all option clear sequence text area. Quit option close your program. Second menu named probe has one option – open reading frames for finding ORFs. Third menu named alignment contains one option – pairwise global for alignment of two genes/DNA regions sequences. And fourth menu named help also contains one option - about program. This option contains short information about program. Open reading frames option of probe menu finds out ORFs, ORFs length, ORFs positions and their final product in the form of amino acid sequence show their result in another output window. Global alignment option of alignment menu opens a new window which contains two sequence boxes, browse button option for genes/DNA regions sequences entry directly or from file, small boxes for changing the value of match, mismatch, and gap. And a SUBMIT button to find the output in a next window. By default value of match, mismatch, and gap is 1, -1, and 2 respectively. Screenshots of this program with analysis of a DNA sequence retrieved from genbank database is given below.

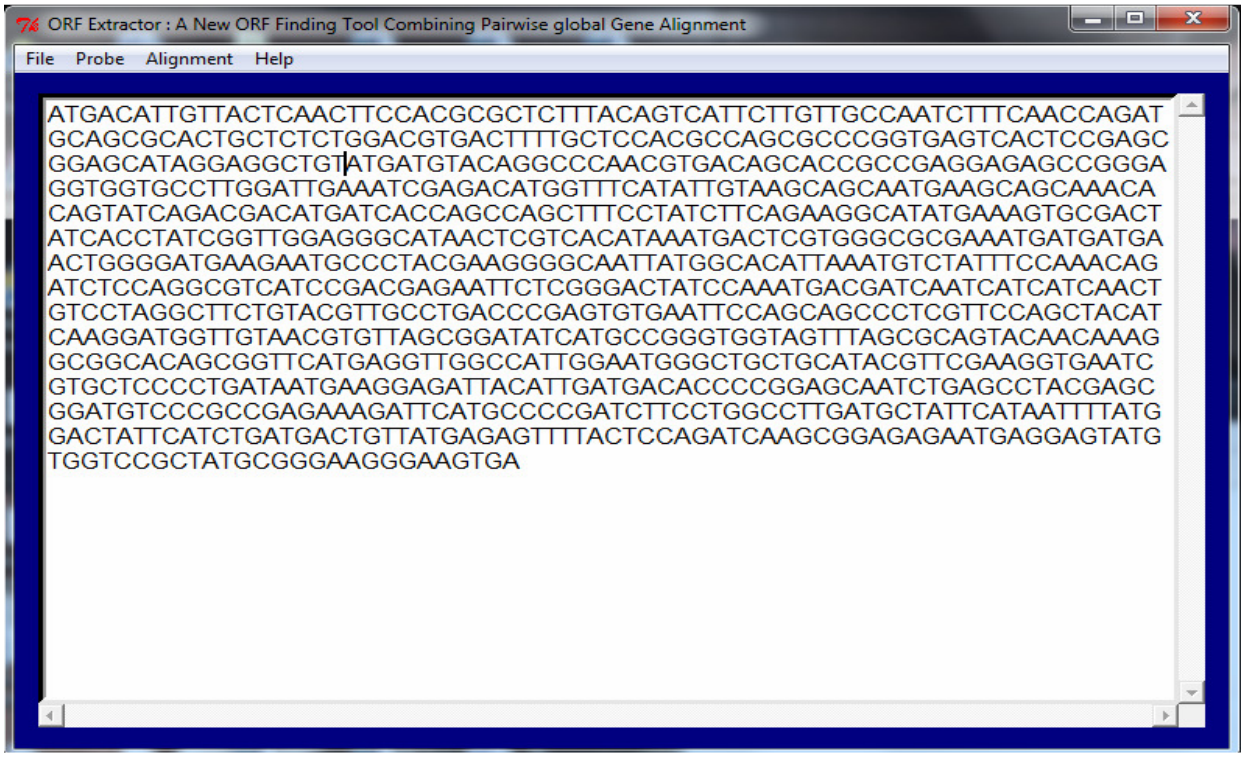

Figure-1
Sequence Input Screenshot

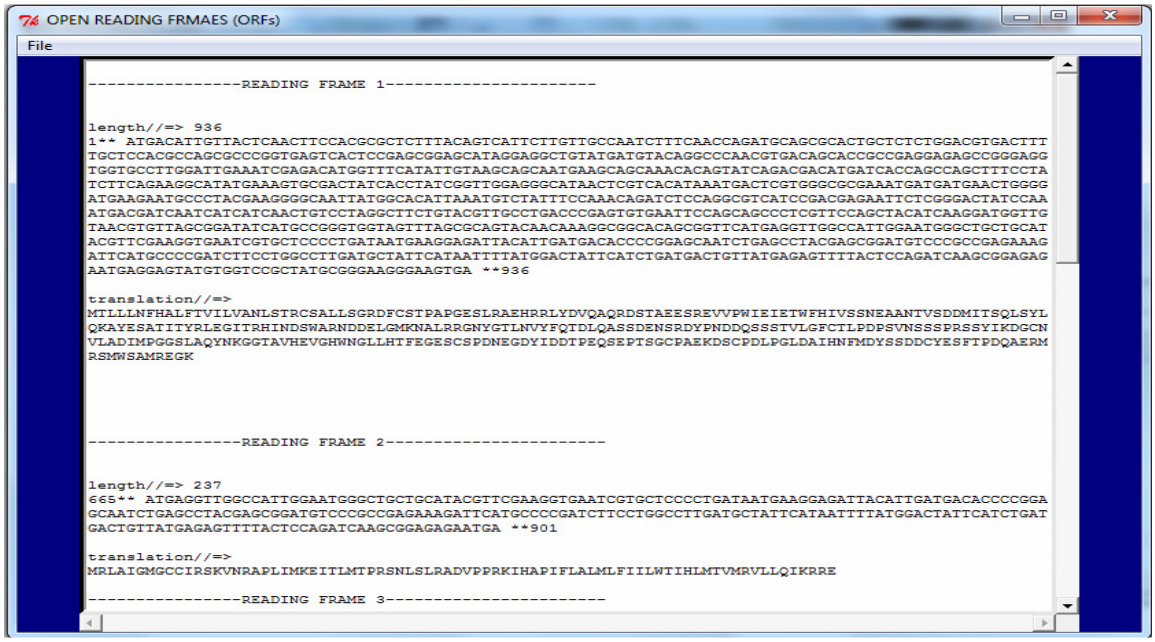

Figure-2
Open Reading Frames Option Output

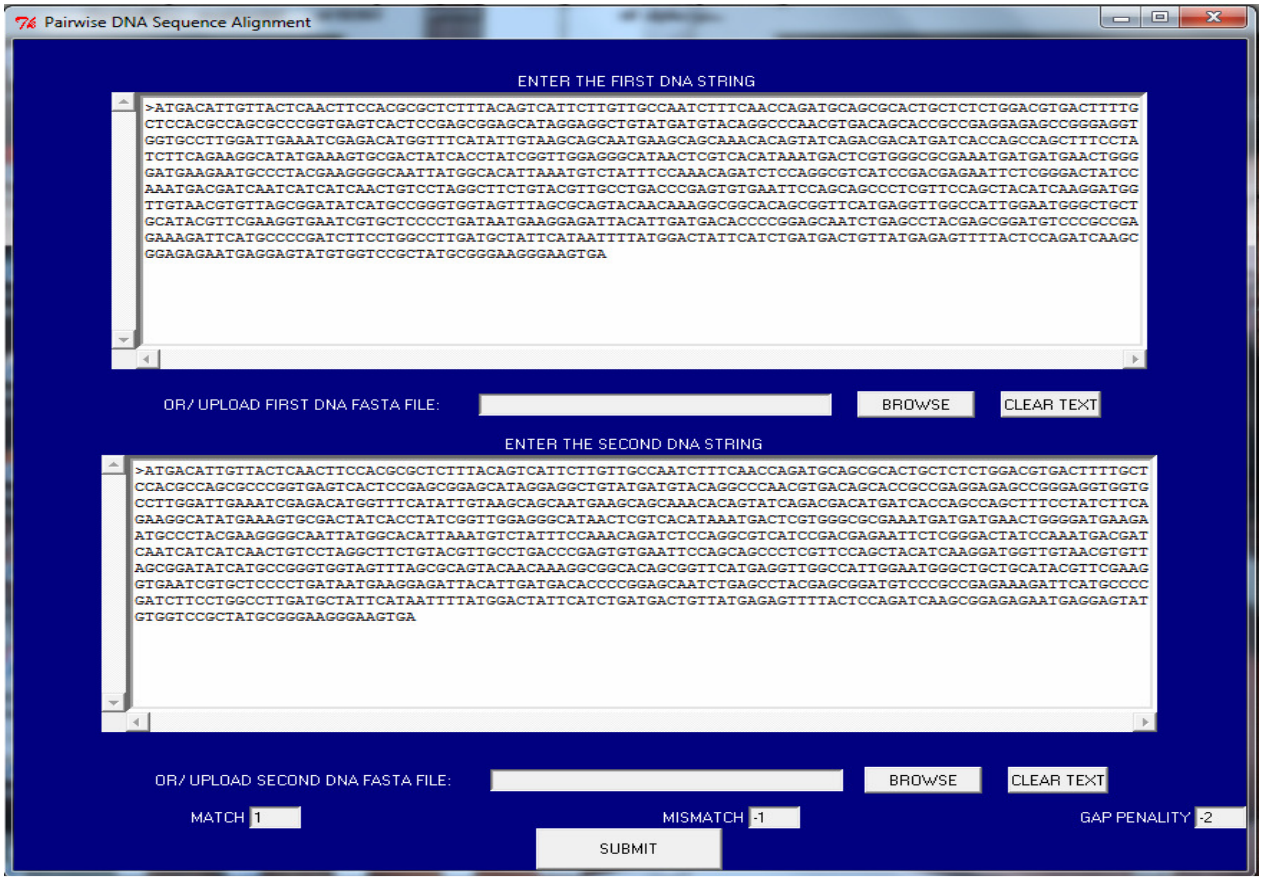

Figure-3
Pairwise Global Alignment Sequence Entry

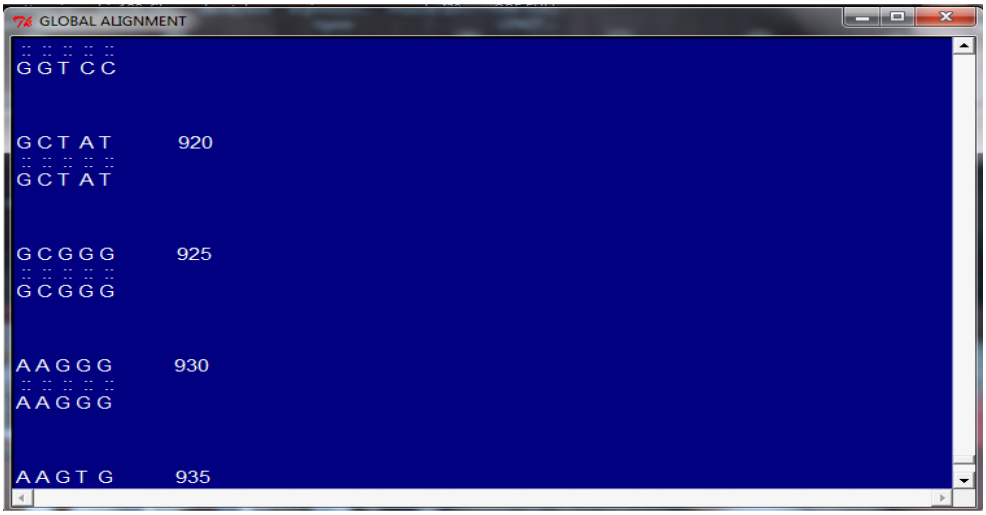

Figure- 4
Output Window of Pairwise Global Gene Alignment

Conclusion

The emerging overall picture is that the sensitivity of ORF Investigator. At present it is not possible to automatically find all genes in a prokaryotic genome. We believe the aim of a gene finding system is to help expert annotators as much as possible, and we consider the statistical significance of a gene an important help in classifying the predictions into almost certain genes and border-line genes needing more attention. With regards to specificity, ORF Investigator performance is very good in comparison to other software tools of gene findings.

Acknowledgements

We are thankful to department of bioinformatics Uttaranchal College of science and technology, Dehradun for providing laboratory facilities and encouragement. I am grateful to my father Mr. Suresh Kumar Dwivedi for his kind support and necessary suggestions whenever I needed.

References

1. Frishman D., Mironov A., Mewes H.W. and Gelfand M., Combining diverse evidence for gene recognition in completely sequenced bacterial genomes, *Nucleic Acids Research*, **26(12)**, 2941-2947 **(1998)**

2. Skovgaard M., Jensen L.J., Brunak S., Ussery D. and Krogh A., On the total number of genes and their length distribution in completemicrobial genomes, *Trends in Genetics*, **17(8)**, 425-428 **(2001)**

3. Kawarabayasi Y., Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix K1. DNA Res*, **6**, 83-101 **(1999)**

4. Fickett J., Recognition of protein coding regions in DNA sequences, *Nucleic Acids Research*, **17**, 5303-5318 **(1982)**

5. Gribskov M., Devereux J. and Burgess R., The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression, *Nucleic Acids Research*, **12**, 539-549 **(1984)**

6. Staden R., Measurements of the effects that coding for a protein has on a DNA sequence and their use in finding genes, *Nucleic Acids Research*, **12**, 551-567 **(1984)**

7. Borodovsky M. and McIninch J., GENMARK: Parallel gene recognition for both DNA strands, *Computers and Chemistry*, **17(2)**, 123-133 **(1993)**

8. Krogh A., Mian I.S. and Haussler D., A hidden Markov model that finds genes in *E. coli* DNA, *Nucleic Acids Research*, **22**, 4768-4778 **(1994)**

9. Salzberg S.L., Delcher A.L., Kasif S. and White O., Microbial gene identification using interpolated Markov models, *Nucleic Acids Research*, **26(2)**, 544-548 **(1998)**

10. Lukashin A.V. and Borodovsky M., GeneMark.hmm: new solutions for gene finding, *Nucleic Acids Research*, **26(4)**, 1107-1115 **(1998)**

11. Besemer J., Lomsadze A. and Borodovsky M., GeneMarkS: a selftraining method for prediction of gene starts in microbial genomes implications for finding sequence motifs in regulatory regions, *Nucleic Acids Research*, **29(12)**, 2607-2618 **(2001)**

12. Besemer J. and Borodovsky M., Heuristic approach to deriving models for gene finding, *Nucleic Acids Research*, **27(19)**, 3911-3920 **(1999)**

13. ORF investigator Program can be downloaded from https://sites.google.com/site/dwivediplanet/ORF-Investigator website.