BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 1 2002
Pages 196–197

Locating probable genes using Fourier transform approach

*Biju Issac, Harpreet Singh, Harpreet Kaur and G. P. S. Raghava**

Institute of Microbial Technology, Chandigarh-160036, India

Received on May 28, 2000; revised on August 17, 2000

ABSTRACT

Summary: FTG is a web server for analyzing nucleotide sequences to predict the genes using Fourier transform techniques. This server implements the existing Fourier transform algorithms for gene prediction and allows the rapid visualization of analysis by output in GIF format.

Availability: The server is available at http://www.imtech.res.in/raghava/ftg/

Contact: raghava@imtech.res.in

Supplementary information: http://www.imtech.res.in/raghava/ftg/help/supl.html

INTRODUCTION

A distinctive feature of the protein-coding genes is the existence of short-range correlations in the nucleotide arrangement, the most prominent being the 1/3 periodicity. Fast Fourier Transformation (FFT) is a powerful technique that is commonly used for detecting periodicity, patterns and tandem repeats in DNA sequences (Fickett and Tung, 1992). Fickett and Tung (1992) describe the significance of Fourier analysis for identification of protein-coding genes in DNA sequence. FFT has the advantage over the other methods like Markov models, neural networks and homology-based methods, of not requiring a learning set, making it possible to identify a protein-coding gene with no known homolog.

GENESCAN, a program developed for gene prediction using FFT, detects the periodicity of three in a nucleotide sequence to locate the probable protein coding region (Tiwari *et al.*, 1997). The FFT is more efficient in detecting the periodicity in a longer sequence ($\geqslant$1024) in comparison to shorter sequences ($\leqslant$150). In order to overcome this limitation of FFT an algorithm called LENGTHEN–SHUFFLE has been described by Yan *et al.* (1998). This algorithm lengthens the short sequences without altering 3-base periodicity. Thus the 3-base periodicity can be predicted with high accuracy using FFT, even in short sequences.

In this Applications Note the authors describe a web server developed for detecting protein-coding genes in DNA sequence using the following Fourier algorithms:

(i) *GENESCAN*. The GENESCAN algorithm considers a sequence of N nucleotides as a symbol string of four symbols **A**, **T**, **G** and **C**. A binary indicator function is defined which selects the elements of the sequence that are equal to the symbol. Four binary sequences are obtained corresponding to each symbol. The Fourier spectrum $S(f)$ is computed from these binary sequences using FFT (Tiwari *et al.*, 1997).
The signal-to-noise ratio of the peak at $f = 1/3$ is computed as $P = S(1/3)/\hat{S}$, where $\hat{S}$ is the average of total spectrum. The sequence or regions in sequence having P greater than or equal to 4 were assigned as protein-coding gene. The plot of f versus $S(f)$ presents the Fourier spectra.

(ii) *LENGTHEN–SHUFFLE*. This algorithm takes a short nucleotide sequence D (e.g. $D \leqslant 150$) and lengthens it K times (where $K = 1200/D$). To eliminate bogus periodicity of D and at the same time keep periodicity of 3 unchanged, the lengthened sequence is shuffled M times (where $M = 10\,000$) with three consecutive bases as a unit. Three different series of digital signals are obtained using the format of *Zcurve*. To these three series of digital signals, FFT algorithm is applied to compute the power spectrum for each series successively called $m1$, $m2$ and $m3$ as described by Yan *et al.* (1998). Threshold T is calculated as, $T = (m1 + m2 + m3)/3$. The sequence or regions having T greater than or equal to 10 were assigned as protein-coding gene (see **Supplementary information**).

(iii) *FTG*. The FTG combines both the above algorithms to improve the accuracy of gene prediction. In the first step it lengthens and shuffles the sequence using the LENTHEN–SHUFFLE algorithm. In the second step it computes the periodicity, Fourier spectra and signal-to-noise ratio using the GENESCAN algorithm.

*To whom correspondence should be addressed.

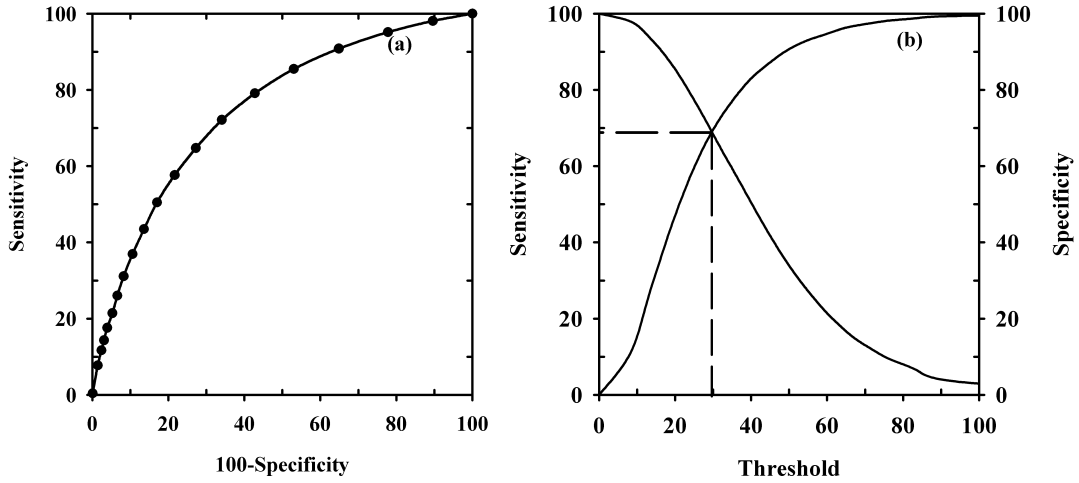

Fig. 1. (a) A plot of 100-specificity versus sensitivity for FTG algorithm on *V.cholerae* chromosome I (window length of 162 and step-size of 5). (b) The plot of threshold versus sensitivity and specificity for FTG (at threshold 30, the sensitivity and specificity are equal).

The performance of FTG-WINDOW was tested on Chromosome I of *Vibrio cholerae* (prokaryote) with a window length of 162 and step-size of 5, and the result is shown in Figure 1a (Heidelberg *et al.*, 2000). The sensitivity and specificity of FTG was ∼69% at threshold 30 (Figure 1b). We also evaluated the performance of FTG in detecting the splice sites, on a dataset of eukaryotic genes (Thanaraj, 2000). The sensitivity and specificity of FTG was ∼61% on eukaryotic genes. It was observed that detection of splice sites using Fourier method(s) alone is difficult and prone to inaccuracy (see **Supplementary information**). Another limitation of FFT is to determine the coding potential of genes lacking 3-base periodicity and discriminating genes from the non-genes that possess 3-base periodicity.

The server uses ReadSeq program (Gilbert, Biology Department, Indiana University) to read the input sequence and can accept most commonly used standard sequence formats. Users can select various parameters such as selecting a particular region in the sequence, limiting window-size, etc. The server analyzes the sequence and presents the result either in text form using HTML or in graphics form using GIF. The server plots the Fourier spectra of a DNA sequence that allows the user to detect any base-length periodicity including 3-base in DNA sequence.

REFERENCES

Fickett,J.W. and Tung,C-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.

Heidelberg,J.F. *et al.* (2000) DNA Sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.

Thanaraj,T.A. (2000) A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures. *Nucleic Acids Res.*, **27**, 2627–2637.

Tiwari,S., Ramachandran,S., Bhattacharya,A., Bhattacharya,S. and Ramaswamy,R. (1997) Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.*, **13**, 263–270.

Yan,M., Lin,Z.-S. and Zhang,C.-T. (1998) A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, **14**, 685–690.