FramePlot: a new implementation of the Frame analysis for predicting protein-coding regions in bacterial DNA with a high G+C content

Jun Ishikawa *, Kunimoto Hotta

Department of Bioactive Molecules, National Institute of Infectious Diseases, Tokyo 162-8640, Japan

Received 8 February 1999; received in revised form 10 March 1999; accepted 10 March 1999

Abstract

FramePlot is a web-based tool for predicting protein-coding regions in bacterial DNA with a high G+C content, such as Streptomyces. The graphical output provides for easy distinction of protein-coding regions from non-coding regions. The plot is a clickable map. Clicking on an ORF provides not only the nucleotide sequence but also its deduced amino acid sequence. These sequences can then be compared to the NCBI sequence database over the Internet. The program is freely available for academic purposes at http://www.nih.go.jp/~jun/cgi-bin/frameplot.pl.

Keywords: High G+C content; Streptomyces

Genes of bacteria which have a high G+C content genome DNA such as Streptomyces have biased codon usage. This results in extremely high G+C distribution at the third letter of each codon. Streptomyces genes actually have an average third-letter G+C content of 92%, which was calculated from 2008 Streptomyces genes recorded in the CUTG database [1] based on the GenBank database release 108. This characteristic enables the prediction of protein-coding region in such bacteria. The Frame analysis was first developed by Bibb et al. [2] and was implemented on a VAX system. Although it is one of the essential analyses for studying Streptomyces genomics at the present time, the software to perform the analysis has been implemented on only a few platforms.

FramePlot is a new implementation of the Frame analysis, with many improvements. The program interface is provided by three web pages: a query sequence submission page, a results page, and a feature page. FramePlot calculates third-letter G+C content within a window of a set ‘window size’ (default 40 codons) and plots the data in the middle of the window. The window is moved along the sequence by a set ‘step size’ (default 5 codons). These running parameters permit a trade-off between speed and resolution of the plot. A higher value of ‘step size’ as well as ‘window size’ yields greater speed, but also a lowered resolution of the plot. In almost all cases, default values will give sufficient results. FramePlot also finds all open reading frames (ORFs) starting...
with a selected ‘start codon(s)’ in the sequence. The ORFs are plotted as bars with potential start (‘>’) and stop (‘|’) codons. The data of each frame is indicated by different colors (color mode) or line-style (black and white mode). Fig. 1 shows the result of the analysis of the sequence containing the kan gene [3]. There is an apparent ORF with extremely high third-letter G+C content at the 434–1288 region.
(Fig. 1A). Clicking on the ORF yields the nucleotide and its deduced amino acid sequence with their frame number, nucleotide position, length, and analysis date (Fig. 1B). Furthermore, the sequences can be compared to the GenBank database over the Internet by using the NCBI BLAST server [4]. This feature helps in finding a new gene.

In the course of the development of the program, we found a new gene. There is a small ORF with high third-letter G+C content at the downstream of the *kan* gene. Deduced amino acid sequence of the small ORF shows homology to hyaluronidase and chondroitinase. Further sequencing study revealed that this small ORF was a 3′ region of a larger ORF. The ORF could encode a 77.3-kDa protein consisting of 721 amino acids with 94.2% G+C in the third position of the codons. Furthermore, the 4.2-kb *Sph*I fragment containing the ORF conferred the ability to grow on hyaluronic acid as a sole carbon source to *S. lividans* TK21. These results would indicate that the gene encodes a hyaluronidase.

*FramePlot* can accept sequence data in any format. In order to analyze low quality sequence data, for example high throughput genomic sequences, all alphabetic characters are acceptable in addition to A, C, G, and T, while non-alphabetic characters, for example digits, spaces, etc., can be removed by the program.

*FramePlot* is freely accessible for academic purposes at http://www.nih.go.jp/~jun/cgi-bin/frameplot.pl. The source code is also available at the same site. To install the program, Perl version 5.0 and Fly version 1.6 or later are required [5]. Commercial users should contact the author for licensing details.

Acknowledgments

We would like to thank Dr. H. Ikeda for thorough tests of the program. We give special thanks to Dr. M.J. Bibb for encouraging comments, Dr. J.A. Gil for critical reading of the manuscript, and Mr. and Mrs. Summers for correcting the manuscript.

References


