A rapid method for illustrating features in both coding and non-coding regions of a genome

R. Hall* and L. Stern

Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia

Received on June 28, 2003; revised on September 18, 2003; accepted on November 7, 2003
Advance Access publication February 5, 2004

**ABSTRACT**

Summary: A bitmap display of the Fourier spectra has been developed which allows convenient whole chromosome scanning for genes and other features. Use of a limited sliding window gives rapid visualization and localization of coding regions in the chromosomes, as well as non-coding features such as repetitive DNA. The method works particularly well on organisms with a skewed base composition, to provide an overview of genomic features.

Availability: Linux Intel and Solaris 8 Sparc binaries available on request.

Contact: rshall@cs.mu.oz.au

Supplementary information: Further Fourier mappings can be viewed at www.student.unimelb.edu.au/~hallrs/

The frequency content of a signal can be determined by the analysis of its Fourier transform (O’Neil, 1991). When DNA is viewed as a signal, a discrete Fourier transform provides a method suited to the detection of periodic arrangements of bases in a genome. Transformation of the base sequence to a numerical array suitable for Fourier analysis can be done in a number of ways (Silverman and Linsker, 1986; Tiwari et al., 1997). Fourier analysis has been used to detect coding regions (Tiwari et al., 1997; Ficket and Tung, 1992) and splice sites (Huestis and Saul, 2001). Numerous software tools for gene-finding are currently available (Burset and Guigo, 1996; Salzberg et al., 1998). However, disagreements between different methods result in uncertainties, and make it desirable to have a spectrum of independent methods available (Miller and Attwood, 2003). This paper shows that an appropriate visual representation of the Fourier analysis can provide a guide to gene location, gene density and other periodic features in the genome. The Fourier method does not require prior knowledge of the genome to be analysed, and its rapid operation suits its application as a first pass analysis.

A numerical representation of a sample DNA is constructed for each of the 4 nucleotides. Four arrays are implemented using binary strings as described by Tiwari et al. (1997), where each occurrence of the relevant base is indicated by a 1 and any other base by a 0. A Fourier transform was performed on each array, and the sum of the squares of the individual spectrum components produced an overall Fourier analysis for the particular sample of the genome. The algorithm was implemented in C.

An overall view of a genomic sequence is constructed by taking sequential, overlapping samples and performing a combined Fourier analysis on each. The spectrum from each analysis is scaled by a constant factor and converted to a column of grey-scale pixels, where high power in the spectrum is represented by a dark pixel. The resulting bitmap has a width of the genome length divided by half the sample size and a depth of half the sample size. Due to the sequential sampling, a small peak occurring at the same location over a number of contiguous samples produces a distinct line in the bitmap. For example, using a sample of 256 bases, a continuous horizontal line at a pixel column height of 85 indicates a strong period 3 in these samples.

Using a sample size of 512 bases, a one million base genome produces a grey-scale bitmap in the pgm format 256 pixels × 3900 pixels in about 20 s on a 333 MHz Sun Ultra Sparc. An X Window program has been developed to allow convenient scrolling of the bitmap image along with scales giving sequence offsets and periodicity values.

The Fourier analysis and visualization was used to examine DNA sequences from a number of organisms. Figure 1 shows the Fourier analyses of the first 100 kb of Plasmodium falciparum chromosome 3 (Bowman et al., 1999) represented as a bitmap, using a sample size of 512 bases. Annotated exons are shown at the top of the figure for comparison. Periodicity is apparent in both coding and non-coding regions.

A well-defined broken line is seen at pixel column height 171. This denotes a strong repeating signal with a period of 3 (512/171 ≈ 3). The period 3 lines correlate well with exon locations as described in Bowman et al. (1999).

Other laddered regions are seen in non-coding regions.

In chromosome 3, a section with a period of 7 is visible in the 500 bases at the beginning of the chromosome. A strong period 12 feature is visible from offset 9.5 to 10.5 kb, followed by a 20 kb section with a prominent period 21 repeat. These features in the first 30 kb of chromosome 3 correspond to...
A rapid method for illustrating genome features

known repeating regions in non-coding DNA. For example, the 20 kb section from 10.7 to 30.7 kb corresponds to the known 21 bp repeat region rep20 (Bowman et al., 1999). Repeating regions have a characteristic 'laddered' appearance due to the harmonics of the Fourier transform (O'Neil, 1991).

When applied to Leishmania major chromosome 1 (62% GC), Streptomyces coelicolor (72% GC) and Escherichia coli (50.4% GC), period 3 lines also showed in coding regions. Due to the high coding density of these genomes, the period 3 lines are almost continuous.

Human coding sequences from the ALLSEQ genefinder benchmarking dataset (Burset and Guigo, 1996) produced few notable results apart from laddered repeat regions on genes such as HUMTRHYAL (trichohyalin).

The Fourier method gives a good indication of coding regions, but its resolution is not fine enough to show exon boundaries clearly even when an incremental sliding window is used. Use of a limited sliding window provides a resolution close to that of the more computationally expensive incremental sliding window and results in a bitmap of large genomic regions that is small enough to provide an overall view.

For visual ease we adjusted the threshold assignment of pixel values, so that a consistent line is seen over the length of the coding region regardless of variation in the strength of the signal. The result is a convenient and rapid overview of a genome. Within a few seconds, the genome of an organism can be assessed for coding density, approximate gene locations and repeat regions. The method is particularly useful on genomes with a biased base composition. Because exon boundaries are not resolved clearly, here or in any other Fourier-based method, the technique is designed to be used as a first-pass, in conjunction with other gene and feature-finding software.

ACKNOWLEDGEMENTS

Sequences for P. falciparum, L. major, S. coelicolor and E. coli were obtained from the National Center for Biotechnology Information (www.ncbi.nlm.gov), National Library of Medicine and National Institutes of Health (USA). An updated P. falciparum annotation was obtained from The Walter and Eliza Hall Institute (http://www.wehi.edu.au/MalDB-www/ae1362ra.htm), Melbourne, Australia.

REFERENCES


