FEBS Letters 581 (2007) 253–258

Noise-reduction filtering for accurate detection of replication termini in bacterial genomes

Kazuharu Arakawa, Rintaro Saito*, Masaru Tomita

Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan

Received 11 September 2006; revised 5 December 2006; accepted 8 December 2006

Available online 18 December 2006

Edited by Takashi Gojobori

Abstract **Bacterial chromosomes are highly polarized in their nucleotide composition through mutational selection related to replication. Using compositional skews such as the GC skew, replication origin and terminus can be predicted in silico by observing the shift points. However, the genome sequence is affected by myriad functional requirements and selection on numerous subgenomic features, and elimination of this "noise" should lead to better predictions. Here, we present a noise-reduction approach that uses low-pass filtering through Fast Fourier transform coupled with cumulative skew graphs. It increases the prediction accuracy of the replication termini compared with previously documented methods based on genomic base composition.**

Keywords: GC skew; Origin of replication; Polarization of chromosome; Fast Fourier transform; Horizontally transferred genes

1. Introduction

Circular bacterial chromosomes are typically highly polarized and segregated into two major regions with asymmetrical preference of complementary nucleotides [1,2]. This compositional bias can be observed using GC skew graphs, which plot the normalized excess of G over C [(C − G)/(C + G)] in sliding windows along the genome sequence [3]. The GC skew graph elucidates two replichores having opposite polarity of GC content that are known to correspond to the two replication arms, with the shift point of GC polarity correlating with the replication origin (*ori*) and terminus (*ter*) located directly opposite each other to maintain a physical balance [4,5]. The leading strand exhibits an excess of G over C, which is attributed to the different mutation and nucleotide-substitution rates and selective pressure on the leading and lagging strands during complementary strand synthesis in replication [6,7]. Several models for this mutational bias in the replichores have been proposed [8], such as T/G mismatch [9] and cytosine deamination [10]. GC strand asymmetry is shown to be the result of global mutational bias, independently of gene orientation, selection for amino acid coding requirements in genes, or skewed signal sequences [11].

To understand the factors controlling evolutionary genome organization through replication, which is central to the bacterial cell cycle, it is essential that we know the locations of *ori* and *ter*, although it is difficult to determine this from direct experiments. However, since the polarized chromosomal features are widely conserved among many bacterial species, in silico methods have been developed to predict the locations of *ori* and *ter*, taking advantage of the availability of complete genome sequences. Unlike with GC skew graphs, the use of DNA walk graphs eliminates the requirement to use sliding windows, and projection of a DNA walk graph onto a GC coordinate system results in a cumulative GC skew diagram that has sharp turning points at the maximum and minimum, corresponding to the locations of *ori* and *ter*, respectively [12]. DNA walk representation can be further enhanced into three-dimensional space by using the Z curve method [13,14]. The use and correlation of AT, keto (GT), and purine (AG), and coding strand excesses, instead of the simple GC skew, has been shown to increase the accuracy of prediction [15]. The use of characteristic short oligomers that are highly skewed has also been proposed for higher precision [16]. These methods are implemented by computational software such as Oriloc [17], and although they have been utilized successfully in many genome projects to predict putative *ori* and *ter* sites, the predicted loci are not always precise [18], especially in the case of the *ter* region [19].

Replication in bacteria is initiated when DnaA proteins bound to DnaA boxes (5′-TTATCCACA-3′ in *Escherichia coli*) near the *oriC* region unwind the DNA duplex for the formation of a replication fork [20]. The DnaA box sequence is conserved among many bacteria with slight changes in the sequence, and the *dnaA* gene is preferentially located around the *ori* region. Therefore, determination of the location of the DnaA box and *dnaA* gene, coupled with a knowledge of genomic strand asymmetry, often leads to more accurate predictions [21]. Similarly, knowledge of the location of two subgenomic features, the RAG motif and *dif* site, enhances *ter* prediction [22]. When an odd number of homologous recombination events occurs during replication, a chromosome dimer is formed that can be lethal unless correctly resolved by XerCD site-specific recombinase at the conserved *dif* sites [23]. Chromosome dimer resolution is effective only when the *dif* site is positioned between long regions of opposite polarity [24], whose locations are identified by the FtsK translocase that travels along the genome recognizing the chromosomal polarity using a short sequence element named the RAG motif

*Corresponding author. Fax: +81 466 47 5099.
E-mail address: rsaito@sfc.keio.ac.jp (R. Saito).

Abbreviations: G, guanine; C, cytosine; A, adenine; T, thymine; FFT, Fast Fourier transform; HGT, horizontal gene transfer

0014-5793/$32.00
doi:10.1016/j.febslet.2006.12.021

(5′-GNGNAGGG-3′ in *E. coli*) [25,26]. The frequency and skew of RAG motifs are strongly selected near the *ter* region in the course of evolution [27]. However, the DnaA box, the *dif* site, and the RAG motifs – and their characteristics – are not completely conserved among diverse bacterial species; therefore these subgenomic features are not suitable for the universal prediction of bacterial *ori* and *ter*.

To predict *ori* and *ter* in bacteria in silico with higher precision using global methods, we present here the application of noise-reduction techniques using fast Fourier transforms (FFTs). Successful utilization of Fourier and wavelet transforms to many computational sequence analyses has been reported elsewhere [28–30]. However, noise-reduction techniques are rarely used for biological problems even though they have been proven effective in the fields of signal processing and analyses. The polarity of bacterial chromosomes is the result of global mutation bias between the two strands during replication, and genomic features not related to this selectional bias such as the codon preferences, can be thought of as "noise". A reduction in the number of subgenomic features that are not related to replicational bias leads to better accuracy in the prediction of replichores.

2. Materials and methods

2.1. Sequences and software

Complete genomes of 382 microbes in GenBank format were obtained from the NCBI RefSeq repository (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Complete listings of the genomes used in this work are available at http://www.g-language.org/data/oriter/. All analyses were conducted using the G-language Genome Analysis Environment version 1.6.2 [31,32].

2.2. Prediction of ori and ter using genomic compositional skews

Three measures of compositional skew were considered in this work: GC skew (C − G)/(C + G), keto excess (A + C − G − T), and purine excess (C + T − G − A), where A, T, G, and C represent the corresponding nucleotide content at a given position. The locations of *ori* and *ter* were predicted by taking the maxima and minima of the cumulative sum of each of the skew indices along the genome sequence at single base-pair resolution. We used a coordinate system for noting the position in the genome, starting at 0 (GenBank starts from 1).

2.3. Noise reduction by elimination of certain genomic regions

Compositional skews may be disrupted by genomic insertions and by lateral gene-transfer events [19]. To take into account the effects of these events as the control for comparison of noise reduction by FFT, we also predicted *ori* and *ter* after masking four types of subgenomic regions, namely, (1) intergenic regions; (2) low/high GC regions; (3) predicted regions of horizontal gene transfer (HGT) based on codon usage; and (4) horizontally transferred genes listed in the horizontal gene transfer database (HGT-DB) [33]. Low/high GC regions were determined by calculating the local GC content using 1000-bp non-overlapping sliding windows along the genome, and by identifying regions where the absolute difference between the average GC content of the genome and the local region was greater than the standard deviation of the GC content of the sliding windows. Putative horizontally transferred genes were predicted by calculating the codon adaptation index (CAI) [34] of each gene using the average codon usage of the genome as a reference set in the calculation of the w-value, and by identifying genes that had CAI values lower than the average CAI by at least one standard deviation.

2.4. Noise reduction by FFT

Fast Fourier transform is an optimized derivation of discrete Fourier transform (DFT) for computational efficiency when the number of sampling units is to the power of two. DFT transforms a given signal in the time domain to values in the frequency domain, representing the frequency components of the input signal. Here the position in the genome was used as the time domain, and the compositional skew was used as the signal. DFT $F(k)$ of a signal of length N, $f(n)$, $n = 0, 1, \ldots, N-1$, at frequency k was calculated as follows:

$$F(k) = \sum_{n=0}^{N-1} f(n) \mathrm{e}^{-\mathrm{i}2\pi kn/N},$$

where $\mathrm{i} = \sqrt{-1}$. The power spectrum $PS(k)$ of $F(k)$ was further defined as

$$PS(k) = |F(k)|^2, \quad k = 0, 1, 2, \ldots, N-1$$

at each frequency k. Because we were interested in the genomic compositional skew with two regions of opposite polarity (i.e., 1 Hz), high-frequency components could be regarded as noise. Therefore, we applied low-pass filtering where the power spectra were zeroed for frequencies larger than a certain threshold, leaving only the low-frequency domain of the signal. The filtered power spectrum was transformed back to the signal data by taking the inverse DFT. The Math::FFT module of Perl (http://search.cpan.org/ ~ rkobes/Math-FFT-1.28/FFT.pm) was utilized for FFT calculation. For calculation efficiency with FFT, genome sequences were divided into 4096 windows, and GC skew and keto/purine excesses were calculated in these windows to obtain the initial signal. Low-pass filtering was conducted by eliminating 10%, 20%, 50%, 90%, 95%, and 99% of the high-frequency regions. Locations of *ori* and *ter* were calculated by taking the maxima and minima of the cumulative sum of the filtered signal.

3. Results

First, to test the applicability of the noise-reduction by FFT, we calculated the power spectrum of the GC skew. The results for the *E. coli* genome are shown in Fig. 1 for all the frequencies analyzed (Fig. 1a) and for the lower-frequency region from 1 to 50 Hz (Fig. 1b). Because the time domain corresponded to the position in the genome in this analysis, 1 Hz denoted one oscillation with two regions of opposite polarity in the genome, thereby correlating to the genomic compositional skew. As expected, the power spectrum was strong at a frequency of 1 Hz but nearly zero at all of the other, higher, frequencies. Therefore, the skew was predominantly affected by selection acting on the replichores, and all other effects were considered as noise for this purpose.

Using this result, a low-pass filter was applied to the GC skew signal, eliminating the spectral components in the high frequency domains. The results for the *E. coli* genome are shown in Fig. 2 with low-pass filters of 0%, 50%, 90%, 95%, and 99% (Fig. 2a–e, respectively). Because the high-frequency components had little effect on the overall skew, low-pass filters up to 90% only gradually reduced the noise. Strong filters at >90% greatly clarified the appearance of the skew graph, reducing the noise of the graph and elucidating the loci of the shift points. The positions of *ori* and *ter* were predicted after low-pass filtering of 0–4096 windows and are shown in Fig. 2f to clarify the effects of the low-pass filtration thresholds on the predicted loci. As with the representations in the skew graphs, low-pass filtering had no effect on the prediction until strong filters above 90% elimination were applied, resulting in infinitesimal changes. Filtration above 99% removed the essential component frequencies and resulted in chaotic predictions.

Finally, *ori* and *ter* of *E. coli* were predicted using GC skew, keto and purine excesses, low-pass filtering by FFT, and elimination of subgenomic regions. The predicted results were compared with experimentally confirmed loci [35] and with