Frequency-Domain Algorithms for Visual Analysis on Genomic Structures in Prokaryotes

Lee Sing CHEONG[1], Feng LIN[2], Hock Soon SEAH[2]
[1]*Bioinformatics Research Centre,* [2]*School of computer Engineering*
Nanyang Technological University, Nanyang Avenue, Singapore 639798
lscheong@pmail.ntu.edu.sg, asflin@ntu.edu.sg, ashsseah@ntu.edu.sg

Abstract

Frequency-domain algorithms are studied to provide the composition of each nucleic acid base simultaneously across numerous frequencies and along the base-positions, through Fourier transformation and color mapping of the bases. And also, composite colored scalograms are constructed by continuous wavelet transformation to provide their individual localized time and periodic energy content in a single figure. The visualization of three-dimensional information in the forms of bases, frequencies and base-positions allows visual analysis of correlated structural and genomic information, which can hardly be obtained from the conventional character string analysis. Significant features of the studied prokaryotic genomes were revealed by the use of the programs developed.

1. Introduction

A genome sequence is defined by a series of the bases adenine (*A*), thymine (*T*), cytosine (*C*) and guanine (*G*). Within the genomic sequence, patterns can occur due to the statistical regularities of the structural features or the repetition of the features. The position of the origin of replication, for example, can be mapped by autoradiography, an *in vitro* method which detects the electrophoretic migration of the replicating restriction fragment [1]. However, as *in vitro* experiments require considerable expertise, time and resources, they are rarely conducted, and instead, *in silico* predictions are proposed.

Previous *in silico* predictions [2-3] used calculation methods such as *GC* skews defined as $(G-C)/(G+C)$, purine skews (*G,A* vs. *C,T*), keto skew (*G,T* vs. *A,C*), Z-curve using a combination of skews, and skewed octamers. They are usually presented as an individual plot, without considering the trend at other periodicity. These methods therefore lack the all-rounded visualization functions at various periodicity, and they are not productive in the whole genome analyses.

Gene identification techniques such as variants of Markov model and Artificial Neural Network were also reported based on known biological features and training data set [4]. But these techniques are limited to well-defined known features, and are unable to provide comprehensive coverage of all features.

In the analysis of genomic sequence, calculation methods using wavelet analysis were used in prediction of the location of biological significant features, such as the isochore [5], CpG islands, exons, introns [6] and nucleosome [7] that are present in eukaryotes. However, previous research was limited to allocation of just a particular individual feature, which does not allow the relationship among the various features to be revealed.

Therefore, we are interested in the visualization of all the patterns within the genomic sequence. This visualization will provide a comprehensive view of all the features, including known biological features and the unknown with potential significance. We propose the use of digital signal processing (DSP) technologies to implement frequency-domain algorithms to visualize the composition of each base simultaneously across numerous frequencies and along the base-positions, for efficient genomic structure analysis, such as identification of patterns suggestive of the gene function category, repetitive nature of subsequences, and other unknown structures with potential significance.

The tool is to handle complex structures exhibited in all genome sequences. It is known that the double helical shape allows DNA to be twisted such that approximately 10.5 bases form a 360-degree turn. In prokaryotes, the helix is further condensed by negative supercoiling [8]. This compact negative supercoiling forms independent domains, each consisting of a loop that is associated with proteins that have not been identified. Furthermore, DNA twist is not perfectly regular. The precise rotation per base pair is not a

Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)
0-7695-2606-3/06 $20.00

constant, resulting in the variation of the width of the major and minor grooves. Their exact conformation (A-form, B-form and Z-form) and variation depend on the base pair present at each position along the double helix, their neighboring base pairs and conformation pressure from supercoiling.

It has also been observed that a small number of tandem repeats can be found in Prokaryotes. Their function within putative genes requires further investigation.

In short, due to the limitation enforced by complex factors - evolution, functional roles, and physical-space structure, the task to simultaneously reveal the structural features and patterns within the genome is extremely challenging. The current microscopy technology is unable to visualize the exact conformation along the DNA. We need to design novel *in silico* methods and processes. For that purpose, we devised the algorithms for visualization of three-dimensional information in the forms of bases, frequencies and base-positions, which allow visual analysis on correlated structural and genomic information. The algorithms are packaged in a tool and implemented as a Matlab program.

2. Visual Analysis with Frequency and Base-position Specific Composition

In the DSP based analysis of a genomic sequence, the visual inspection of its spectrogram can provide many distinct visual patterns that correspond to structures and features of the inspected genome. The genomic sequence is first converted into four indicator sequences $\{a(n), t(n), c(n), g(n)\}$ where the value at each location n of the indicator sequence represents the presence (value of 1) or absence (value of 0) of its respective base A, T, C and G at the corresponding position in the genomic sequence.

We display the magnitude of the frequency and base-position specific composition for each of the four bases A, T, C and G simultaneously by superpositioning their corresponding intensity onto a Cartesian plane, using the respective colors blue, red, green and yellow. The grayscale intensity of this color code correlates with the Giemsa staining, in which *AT*-rich regions produce dark regions and *CG*-rich regions produce light regions.

The RGB color space representation of the four colors represents frequency and base-position specific composition. They are mapped using the following equation:

$$X^{color}(f,k) = \Sigma_{B=A,T,C,G} \, (c^{color}{}_{base}) * B(f,k) \,, \qquad (1)$$

where f is the frequency, k is the window index, $c^{color}{}_{base}$ are the RGB color mapping coefficients, and $B(f,k)=\{A(f,k), T(f,k), C(f,k), G(f,k)\}$ are the frequency and base-position specific composition. In this way, we get the RGB color space representation $X^{color}(f,k)=\{X^{Red}(f,k), X^{Green}(f,k), X^{Blue}(f,k)\}$.

The frequency and base-position specific composition $A(f,k)$, $T(f,k)$, $C(f,k)$, $G(f,k)$ can be obtained by short-time Fourier transformation of their respective indicator sequences $a(n)$, $t(n)$, $c(n)$ and $g(n)$, using the following Fourier transformation equation:

$$B(f,k) = \Sigma_{n=base\text{-}positions\ in\ k\ window} \, b(n) \, e^{(-j2\pi fn/N)} \,, \qquad (2)$$

where n is the base position, N is the window length, $B(f,k)=\{A(f,k),T(f,k),C(f,k),G(f,k)\}$ represents the frequency and base-position specific composition of the respective transformed indicator sequence and $b(n)=\{a(n),t(n),c(n),g(n)\}$ represents the indicator sequence of the bases.

Two optional graphs provide the average magnitude of the composition for each of the bases along the frequencies and along the base-positions respectively. The average magnitude of the composition along the frequencies is obtained by averaging the magnitude at all base-positions using the following equation:

$$B(f) = \Sigma^{all\ values\ of\ k} \, |B(f,k)| \,/\, n_w \,, \qquad (3)$$

where n_w is the number of windows and $B(f)=\{A(f),T(f),C(f),G(f)\}$ represents the average magnitude of the composition for the respective base at frequency f and window index k.

The average magnitude of the composition along the base-positions is obtained by averaging the magnitude at all the frequencies except frequency zero using the following equation:

$$B(k) = \Sigma^{all\ values\ of\ f,\ where\ f \neq 0} \, |B(f,k)| \,/\, (n_f - 1) \,, \qquad (4)$$

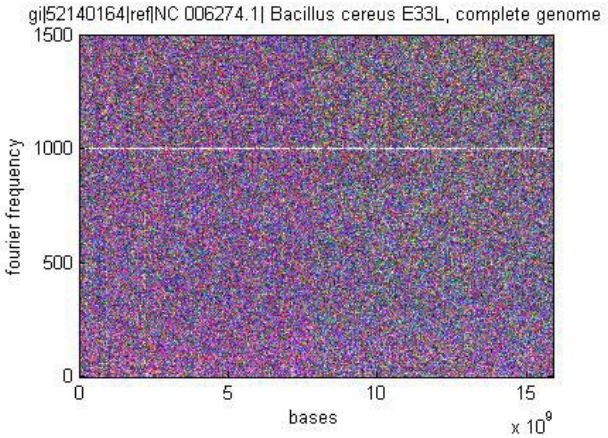
where n_f is the number of frequencies and $B(k)=\{A(k),T(k),C(k),G(k)\}$ represents the average of the magnitudes at all the frequencies except frequency zero for the k^{th} window.

2.1. Origin of Replication

We applied the visualization algorithms discussed in the last section to the investigation of the bacteria genome sequences and obtained interesting findings.

As illustrated in Figure 1, it was found that 12 bacteria (7 unique species) from the *Bacillus* family show two regions of different composition, which are represented by higher intensity of reddish-purple for

one region and a higher bluish-purple for the other region. The two regions of different colors extend from low frequencies till the maximum frequency investigated, including the frequency corresponding to period-3. The base-composition of these two regions is asymmetry for base *A* with *T* and base *C* with *G*. The reddish-purple region has higher intensity of base *T* over base *A* and higher intensity of base *G* over base *C*.



(a) Frequency and base-position specific composition

(b) Average magnitude of the composition

Figure 1: Display of the base-composition asymmetry in the *B. cereus E33L (ZK)* genome

The *GC* base-composition asymmetry correlates with the strand direction of the replication forks in bacteria. The leading strand has a preference of base *G* over base *C*, while the lagging strand has a preference of base *C* over base *G*. The differences in the synthesis of the two strands, where the leading strand is being synthesized continuously, while the lagging strand is being synthesized discontinuously by the use of Okazak fragments, cause bias pressure which lead to the *GC* base-composition asymmetry [9-10].

The *AT* base-composition asymmetry correlates with the type of polymerase-α subunit that replicates the leading strand in different bacteria. The intersection points of the *GC* base-composition asymmetry or the *AT* base-composition asymmetry indicate the position of the origin and terminus of replication.

It was also observed that 4 bacteria (3 unique species) from the *Chlamydia* family show different *GC* base-composition asymmetry and at a lesser degree, *AT* base-composition asymmetry.

This variation of strength in base-composition asymmetry depends on the variation of bacteria and the frequencies under inspection. Thus, the visualization of the base-composition asymmetry at all different frequencies, allows the bias to be extracted at the most suitable frequencies.

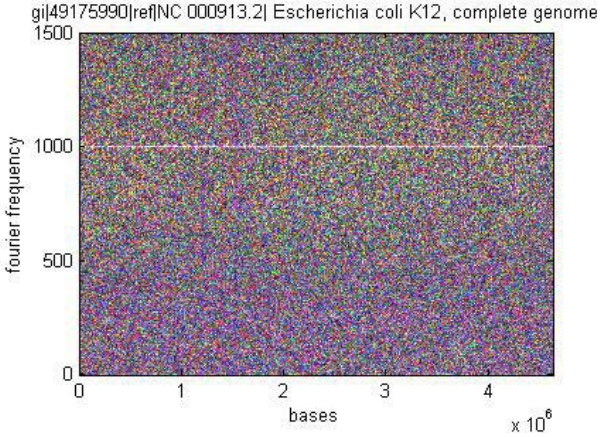
2.2. Patterns of Structural Significance

Our next experiment focuses on changes of frequency throughout a whole genome sequence. In structural genomics study, bacterium genome is characterized as compact clump or series of clumps, condensed by unidentified proteins. It is suggested that the genome is supercoiled into many independent domains, with each domain consisting of a loop secured by the unidentified proteins. Although the genome structure is organized into definite packages, it does not display the eukaryotic chromosomal distinctive morphological features. Thus, it is interesting to find feature which aid in the understanding of the organization of the bacteria genome.

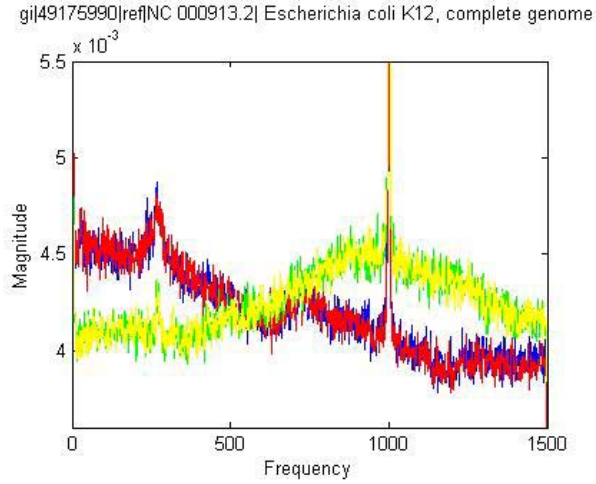
10 bacteria (6 unique species) were placed under close examination. As illustrated in Figure 2, we observed the base-composition pattern consisting of a shift in color from green at higher frequency to purple at low frequency throughout their genome sequences. The green color represents region rich in bases G and C, while the purple color represents region rich in bases A and T. This shift in the color occurs throughout the entire genome, indicating that the characteristic is prevalent genome-wide.

The intersection frequency of the change in the composition corresponds to around period-6. It is noted that period-6 corresponds approximately to a 180 degree turn of the Z-form double helix, and period-6 is also twice the length of a codon. As the characteristic is prevalent genome-wide, this pattern might influence the chromosomal structure, affecting the DNA double-helix conformation and the supercoiling. The exact conformation along the DNA is extremely important, as the width of the major and minor grooves can influence the attachment of binding factors to the DNA.

It is noted that five of the six unique species belongs to *Proteobacteria*, and one of the species belong to *Chlorobi*. Thus, study of this phenomenon is not limited to *Proteobacteria*.



(a) Frequency and base-position specific composition



(b) Average magnitude of the composition

Figure 2: Structural significant pattern in the *E. coli K12-MG1655* genome

2.3. Adjacent Repetitive Sequences

A distinct pattern which represents adjacent repetitive sequences can be observed. This pattern is shown as highly similar columns.

A scan through the *E. coli O157:H7 EDL933* complete genome shows a number of such repeats. Some of the repeats are found within Open Reading Frame (ORF), which are putative proteins, while others partially span or encompass ORF and the adjacent intergenic regions. One repeat found in the ORF is shown in Figure 3.

In the figure, the tandem repeats consisting of approximately 500 nucleotides is part of the putative membrane protein (primary locus: Z2636, from 2,381,838-2,384,060 nucleotide), located near the 3' end.

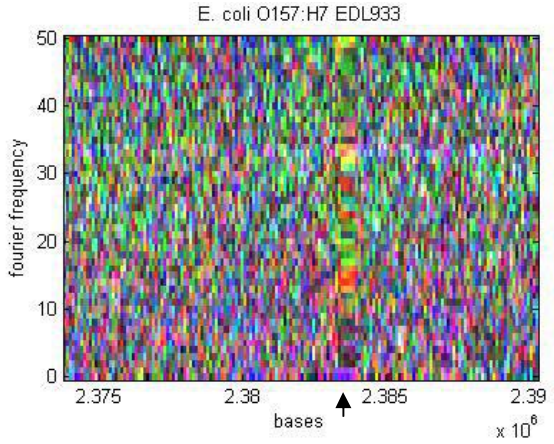


Figure 3: Repeat at 2383601 – 2384100 nucleotide in *E. coli O157:H7 EDL933*

Intergenic tandem repeats are associated with regulatory mechanism, and as a source of genome instability, they can be used in genomic fingerprinting of different strains. A tandem repeat found within putative protein may have resulted in either a gain or loss of gene function of the protein through frameshift during the formation of the repeat [11].

3. Wavelet Analysis

Wavelet analysis is a multiresolution analysis method, allowing both the frequency and time localization information to be simultaneously extracted from an aperiodic signal. It provides information that is different from the Fourier analysis which provides the total spectrum content.

Starting with the indicator sequences $\{a(n), t(n), c(n), g(n)\}$, we apply continuous wavelet transformation to each of them, and their results are combined into a composite colored scalogram in order to provide an additional dimension on top of the traditional pseudo-colored scalogram.

The localized time and periodic energy content $A(j,k)$, $T(j,k)$, $C(j,k)$, $G(j,k)$ can be obtained by continuous wavelet transformation of their respective indicator sequences $a(n)$, $t(n)$, $c(n)$ and $g(n)$.

By using the Morlet wavelet $\psi(n)$ as the analyzing wavelet, we have the following transformation equation:

$$B(j,k) = |\Sigma_n\, b(n)\, \psi_{j,k}(n)|^2, \qquad (5)$$

where n is the base position, $\psi_{j,k}(n) = (1/\sqrt{j})\,\psi((n-j)/k)$ are the dilated and translated wavelet functions of the analyzing wavelet, and $b(n)=\{a(n),t(n),c(n),g(n)\}$ represents the indicator sequence of the bases. In this

way, we obtain $B(j,k)=\{A(j,k),T(j,k),C(j,k),G(j,k)\}$ representing the localized time and periodic energy content of the respective transformed indicator sequence.

The localized time and periodic energy content $B(j,k)$ tends to show high intensity, which is correspondingly shown as a superposition of the base-composition color in the image of constructed scalogram, when the stretch of sequence of length j, with position centralized at location k, reasonably coincide with the dilated and translated wavelet function $\psi_{j,k}(n)$. The recognizable patterns of the high intensity are therefore a visual indicator of a structure found. For example, the pattern of similar colors at regular interval is indicative of regions containing repeats. Depending on the number and length of the repeating units, repeats can be categorized into simple sequence repeats, micro satellites, mini satellites and repetitive motifs.

The wavelet analysis can be applied to the investigation of the smallest known autonomous replicating genetic units - viroids. We use the developed composite color scalogram analysis programs for the genome-wide similarity comparison. The investigation on host interaction based on the understanding of possible structural conformations of viroid during interaction, with insights provided by their scalogram, will provide further understanding on how minimal RNAs can elicit disease.

Using our visualization tool, all the 36 completed sequenced species of viroids that are currently available at the NCBI database are examined. In their genome scalograms constructed by continuous wavelet transformation, we find that each of the viroids contains high base-position localized energy content at regular base positions and at different scale length, as illustrated in Figure 4. This pattern corresponds to the biological structure of sequence repeats generated by sequence duplication and deletions [12].

It is interesting that each of the viroids also displays distinctive curve-like high energy content at the lower scales. This profile may be related to the difference in their specific sequences, and may have potential significance. It is also observed that some viroids share similar localized energy content at various scale level. The similar localized energy content at lower scale level may correspond to the result of recombination among the different species. It provides us with important clues to the evolution of the different species.

In the structural genomics, the composite color spectrogram of these viroids plays a key role in elucidation of the possible conformations and interaction it may assume with the host.

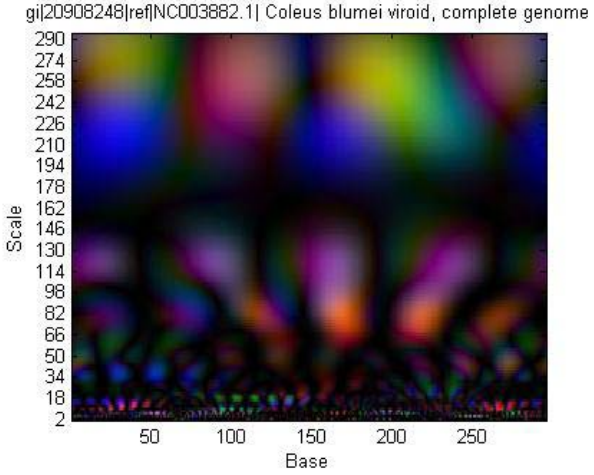


Figure 4: Composite Color Scalogram of the *Coleus Blumei* viroid genome

Diseases are also caused by other pathogenic agents, such as virus and bacteria. In our experiments with *E. coli*, it is noticed that many distinctive features can be identified on the color composite spectrogram.

4. Discussions and Conclusions

The proposed frequency-domain algorithms provide visualization of whole prokaryotic genome and allow structural and genomic analysis. The prokaryotes have been shown that their base-composition asymmetries can be visualized at various frequencies. The intersection of the asymmetry can be used to identify the origin and terminus of replication, which have various applications in genomic study. Identification of the origin of replication provides a landmark to align and locate organizational features using methods such as gene order, thus narrowing the amount of sequence to search. It is also known that most active transcription units are orientated in the same direction as the replication fork, and elements such as DnaA boxes are located near the origin. The origin of replication and its associated trans-acting factors may also be used as drug targets, as competitive binding to them may block the process of replication by depriving access to the origin site and preventing the trans-acting factors from binding to the origin.

Our preliminary study using the tool shows that base-composition asymmetry might also be observed in other bacteria families. In fact, this characteristic has also been reported in other literatures [2-3] as a general feature among bacteria with circular chromosome, and the tool we developed will be greatly helpful in the allocation of the position of origin and terminus of replication and other applications.

The revealed base-composition shift in *Proteobacteria* and a *Chlorobi* is prevalent genome-wide. It helps us understand how the exact conformation along the DNA and the supercoiling of the chromosome is influenced.

Several tandem repeats have been successfully found in *E. Coli*, either within ORF or encompassing ORF and adjacent intergenic regions. The tandem repeats found within putative protein may have resulted in either a gain or loss of gene function of the protein through frameshift during the formation of the tandem repeat [11].

The composite color scalogram allows genomic information to be visualized. The viroids have been shown that their repetitive structures can be visualized as high base-position localized energy content at regular base positions and at different scale length. This pattern corresponds to the biological structure of sequence repeats generated by sequence duplication and deletions. Similar patterns among the viroids are also observed, and may correspond to the recombination among the different species. Distinct patterns at lower scale, is indicative of their difference at the sequence level.

In conclusion, frequency-domain algorithms are studied to provide the composition of each nucleic acid base simultaneously across numerous frequencies and along the base-positions, through Fourier transformation and color mapping of the bases. And also, composite colored scalograms are constructed by continuous wavelet transformation to provide their individual localized time and periodic energy content in a single figure. The visualization of three-dimensional information in the forms of bases, frequencies and base-positions allows visual analysis of correlated structural and genomic information, which can hardly be obtained from the conventional character string analysis. Significant features of the studied prokaryotic genomes have been revealed by the use of the tool.

References

[1] Benjamin Lewin, “The Replicon”, Genes VIII, Pearson Education International, New Jersey, 2004.

[2] Peder Worning, Lars J. Jensen, Peter F. Hallin, Hans-Henrik Staerfeldt and David W. Ussery, Origin of Replication in Circular Prokaryotic Chromosomes, Environmental Microbiology, v.8(2), pp. 353-361, 2006.

[3] Pawel Mackiewicz, Jolanta Zakrzewska-Czerwinska, Anna Zawilak, Miroslaw R. Dudek and Stanislaw Cebrat, Where does Bacterial Replication Start? Rules for Predicting the oriC Region, Nucleic Acids Research, v.32(13), pp. 3781-3791, July 2004.

[4] Catherine Mathe, Marie-France Sagot, Thomas Schiex and Pierre Rouze, Current Methods of Gene Prediction, their Strengths and Weaknesses, Nucleic Acids Research, v.30(19), pp. 4103-4117, August 2002.

[5] Lio P., Vannucci M., Finding pathogenicity islands and gene transfer events in genome data, Bioinformatics, v.16(10), pp. 932-940, October 2000.

[6] Arnéodo A., D'Aubenton-Carafa Y., Audit B., Bacry E., Muzy J.F., Thermes C., Nucleotide composition effects on the long-range correlations in human genes, The European Physical Journal B - Condensed Matter, v.1(2), pp. 259-263, February 1998.

[7] Audit B., Vaillant C., Arneodo A., D'Aubenton-Carafa Y., Thermes C., Wavelet analysis of DNA bending profiles reveals structural constraints on the evolution of genomic sequences, Journal of Biological Physics, vol. 30, no. 1, pp.33-81, 2004.

[8] James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine, Richard Losick, "The Structures of DNA and RNA", Molecular Biology of the Gene, Benjamin Cummings, San Francisco, 2003.

[9] Eduardo P.C. Rocha, The Replication-Related Organization of Bacterial Genomes, Microbiology, v.150, pp. 1609-1627, 2004.

[10] Pilar M. Francino and Howard Ochman, Deamination as the Basis of Strand-Asymmetric Evolution in Transcribed Escherichia coli Sequences, Molecular Biology and Evolution, v.18(6), pp. 1147-1150, June 2001.

[11] You-Chun Li, Abraham B. Korol, Tzion Fahima and Eviatar Nevo, Microsatellites Within Genes: Structure, Function, and Evolution, Molecular Biology and Evolution, v.21(6), pp. 991-1007, February 2004.

[12] Ricardo Flores, Randles J.W., Bar-Joseph M., Diener T.O., A Proposed Scheme for Viroid Classification and Nomenclature, Archives of Virology, v.143(3), pp. 623-629, 1998.