The 47th IEEE International Midwest
Symposium on Circuits and Systems

Fourier and Spectral Envelope Analysis of Medically Important Bacterial and Fungal Sequences

Katie K. H. Chan, Chunqi Chang, and Francis H. Y. Chan

Department of Electrical and Electronic Engineering, The University of Hong Kong,

Pokfulam Road, Hong Kong

***Abstract*-In this paper, we introduce the Fourier and spectral envelope analysis methods to analyze some biomolecular sequences, particularly medically important bacteria and fungi DNA sequences, to get their interesting frequency properties. Fourier analysis includes mapping character strings into numerical sequences, calculating spectra of DNA sequences and setting and solving optimization problem in order to construct a powerful predictor of exons along the long DNA sequences. The spectral envelope analysis makes use of spectral envelope for analyzing periodicities in categorical-valued time series and it is useful for the scaling of non-numeric sequences. The spectral envelope analysis utilizes optimization procedure to improve upon traditional analysis performance in distinguishing coding from non-coding regions in DNA sequences. The two approaches greatly facilitate the understanding of local nature, structure and function of biomolecular sequences. They also provide useful techniques to combine bioinformatics analysis with modern computer power to quickly search for diagnostic patterns within long sequences.**

Key words: Frequency analysis, Fourier analysis, Spectral envelope analysis, DNA sequences

I. Introduction

Recognizing the coding regions within a DNA sequences is a complicated task. Fourier techniques can be used to analyze the biomolecular sequences and predict the locations of exons (protein coding regions),at where nucleotides can be used in the protein synthesis, within a DNA stretch. Cell functions can then be revealed by investigating the proteomics information [1-3].

Genomic information is digital in a real sense, it comes in sequences of character strings where each element is one out of a finite number of possible entities [4].

The spectral envelope approach is a statistical basis on which to establish the analysis of categorical time series in the frequency domain. One approach for investigating the periodic nature of a categorical process is to assign numerical values to each of the states or categories followed by a spectral analysis of the resulting discrete-valued time series. The spectral envelope is an extension of spectral analysis when the data are categorical-valued such as DNA sequences [5].

The purpose of our study was to compare these two analyzes and test their capabilities in detection of coding regions of long bacterial and fungal sequences. Many biomedical applications make use of Fourier-transform-based approaches to analyze signals. Fourier transform and spectral envelope analysis are two of them [6-8].

II. Methods

The short-time Fourier transform (STFT) and nonparametric frequency estimation were tested using the same high-resolution bacteria and fungi DNA data. The Fourier analysis can refer to D. Anastassiou's research [9].

A. Spectral Envelope Analysis

The FFT spectra with different sizes of window for around 10000 nucleotides data segments were calculated by periodogram method [10].

Similar to the numerical assignment shown in the previous approach, the spectral envelope , $\lambda(\omega)$, can be found by first forming 3x1vectors:

$$Y_t = (1,0,0)' \text{ if } X_t = A$$

$$Y_t = (0,1,0)' \text{ if } X_t = C$$

$$Y_t = (0,0,1)' \text{ if } X_t = G$$

$$Y_t = (0,0,0)' \text{ if } X_t = T$$

The algorithm shown below was used to estimate the spectral envelop of a certain size, which is one of the parameters of the test, for a particular part of the sample sequence:

1. With the sample sequence of length N, obtain the 3 x 1 vectors Y_t, where t = 1,2,...,N.
2. Calculate the Fast Fourier Transform(FFT)

0-7803-8346-X/04/$20.00

of the data:

$$d(k/N) = k^{-1/2}\sum_{t=1}^{n} Y_t e^{-2\pi\pi tj/} \quad (1)$$

Notice that d(k/N) is a 3 x 1 complex-valued vector.

3. Calculate the periodogram, f(k/N) = d(k/N)d*(k/N), for k = 1,2,…,(N/2).
4. Perform smoothing which can be done by

$$\hat{I}^{re}(k/N) = \sum_{q=-m}^{m} h_q f^{re}([k+p]/N) \quad (2)$$

where h_q is a parameter for determining the type of window used for smoothing the periodogram. The degree of smoothness can also be controlled.

5. Calculate the 3 x 3 variance-covariance matrix of the data.

$$C = N^{-1}\sum_{t=1}^{n}(Y_t - \overline{Y})(Y_t - \overline{Y})' \quad (3)$$

where $\overline{Y}$ is the sample mean of the data.

6. For each of ω_k = k/N, find the largest eigenvalue and the corresponding eigenvector of the matrix

$$P = 2N^{-1}C^{-1/2}I^{re}(\omega_k)\,C^{-1/2} \quad (4)$$

Notice that $C^{1/2}$ is the unique square root matrix of C and $C^{-1/2}$ is the inverse of that matrix.

7. The sample spectral envelope $\lambda(\omega)$ is the eigenvalue obtained in the above procedures.

Since it was observed that there is frequency 1/3 signal whenever the spectral envelope lies within a coding region, the value of this particular frequency can be collected along a longer segment of the bacterial sequence for predicting exon locations. For instance, if the size of analyzing window is 126, the value at 43th position is obtained since 126/3 +1 = 43. With any modern and sophisticated programming language, the fast Fourier Transforms, eigenvalues and eigenvectors of real symmetric matrices can be computed efficiently.

III. Results

The spectral envelope was made to slide across the sequence to detect the protein coding regions. If the envelope lies within the coding region, it would give high peak at its W/3+1 frequency, where W is the size of the envelope. This well-known phenomenon was tested with the analysis approach mentioned in previous part [5]. A genomic signal processing platform was built using Matlab.

Here, we are going to show the result from our analysis for fungal sequence segments. A DNA stretch, Schizosaccharomyces pombe (GenBank accession number AL136078) was examined. Schizosaccharomyces octosporus (GenBank accession number NC_004312) is used as its reference sequence. The tested DNA stretch consists of 8000 nucleotides beginning from location 21001 through 29000.

Table 1
Exon locations of Schizosaccharomyces pombe

Relative Location	Gene Length	Forward or Reverse Complement
21060→22556	1497	Reverse Complement
24301→25554	1254	Reverse Complement
26337→26642	306	Forward
27150→28457	1308	Forward

There are 4 protein coding regions within this stretch, which are shown below:

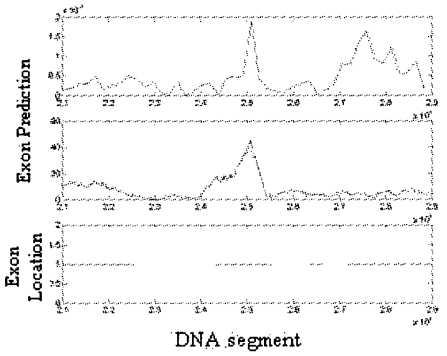

Fig. 1: Exon Predictor for a DNA stretch from Schizosaccharomyces pombe

In Fig. 1, the upper trace shows the result obtained from optimal Fourier analysis and the middle trace shows the result of the spectral envelope analysis. The bottom trace shows the exact locations of the coding regions. All the coding regions are annotated in the GenBank sequence file [11]. The values of parameters are shown as follows:

Table 2
Parameter table of sequence analysis on Schizosaccharomyces pombe

Data Window(w)	Overlap region of data window(r)	Type of Smoothing Window	Smoothing Window Size(m)
351	176	Triangular	4

Another fungal example is shown by using Saccharomyces cerevisiae (GenBank accession number NC_001147). The tested DNA stretch consists of 10000 nucleotides beginning from location 15001 through 25000.

Table 3
Exon locations of Schizosaccharomyces cerevisiae

Relative Location	Gene Length	Forward or Reverse Complement
15232→15504	273	Reverse Complement
17280→17795	516	Reverse Complement
19490→21310	1821	Reverse Complement
22524→24293	1770	Reverse Complement

Within this segment stretch, there are 4 protein coding regions. They are shown in the following diagram:

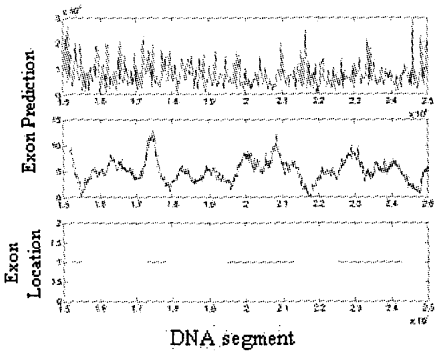

Fig. 2: Exon Predictor for a DNA stretch from Saccharomyces cerevisiae chromosome XV

In Fig. 2, the upper trace shows the result obtained from traditional non-optimal Fourier analysis and the middle trace shows the result obtained from the spectral envelope analysis. The bottom trace shows the exact locations of the coding regions. The values of parameters implemented are shown as follows:

Table 4
Parameter table of Spectral Envelope Analysis on Saccharomyces cerevisiae

Data Window(w)	Overlap region of data window(r)	Type of Smoothing Window	Smoothing Window Size(m)
702	351	Triangular	4

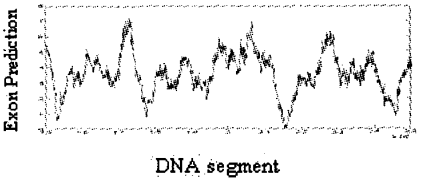

Figure 3: Exon Predictor for Saccharomyces cerevisiae with w=702, m=10

It is observed that larger data window size gives better peak visualization of coding regions for the sequence. It is due to the fact that lengths of some exons within the segment are quite large in size and they are relatively sparsely spaced. Thus, a relatively large data window can give better analysis of the sequence. However, in this case, increasing the smoothing window does not give a better view for the presence of coding regions. Also, notice that some of the peaks become less distinctive because of the enlarged size of smoothing window, compared with the middle trace from Fig. 2.

By comparing the traditional non-optimal Fourier analysis and spectral envelope analysis:

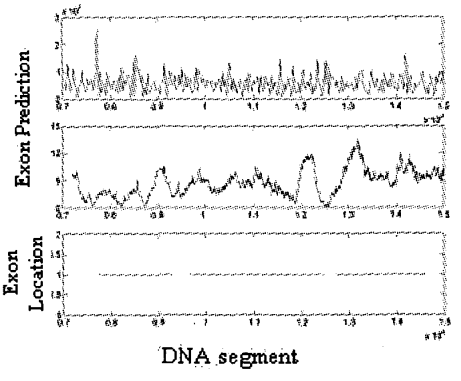

Figure 4: Exon Predictor for Haemophilus phage PH2

In the above diagram, the top trace shows result of traditional non-optimal Fourier method, it is measured by the frequency spectrum $S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 +$

$|U_G[k]|^2$. The middle and the lowest trace show the exon predictor by spectral envelope analysis and the actual exon locations respectively. It is clearly shown that the signals of protein coding regions are much stronger in the Spectral Envelope Analysis; whereas signal peaks of non-optimal method do not demonstrate much difference between introns and exons.

IV. Discussion and Conclusion

From our empirical observation, the results of digital signal processing analysis for bacterial sequences are not very promising, i.e. the existence of exons are not very clearly shown. The test was done on several bacterial sequences including Haemophilus phage PH2 (Accession No. NC_003315) [12]. It is due to the fact that Bacteria are prokaryotes, their inter-coding area in their genomes are always very short. Therefore, DSP methods (both Fourier and spectral envelope analysis) are not the appropriate tools for detect their protein coding regions. Hence, bacterial sequences are suggested to be analyzed by some other mathematical models such as wavelet analysis. On the other hand, the DSP analysis seems a good tool for examining fungal sequences.

The Fourier analysis and the spectral envelope analysis showed consistency in indicating the locations of protein coding regions, where exons lie. The display of those regions was greatly depended on the parameters: data window size(w), overlapping regions of data window(r), smoothing window size(m) and also the type of the smoothing window. Apart from the importance of parameters, for the spectral envelope analysis, there is also a close relationship between envelope size and the size of protein coding region. If the protein coding regions are close together, a small spectral envelope is more preferable for exons detection than a large window. All these values should be carefully chosen for better indication for those regions so that the study of the medically important bacteria can be greatly facilitated.

Obviously, spectral envelope analysis is a more convenient method for studying the protein coding region since only the tested sequence is needed in this analysis; whereas a reference sequence is also required for the Fourier analysis.

Acknowledgements

This work was supported in part by the Research Grants Council of Hong Kong under Grant HKU7180/03E. The author K. K. H. Chan would like to express her profound gratitude to her supervisor at the Microbiology Research Laboratory in the Faculty of Medicine, The University of Hong Kong, Dr. Susanna Lau, and all the co-workers for their ever-ready support and guidance.

References

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walkter, *Essential Cell Biology: The Introduction to the Molecular Biology of the Cell*, Garland Publishing, New York, 1/e,1997.

[2] Anthony J.F. Griffiths, William M. Gelbart, Jeffrey H. Miller, and Richard C. Lewontin, *Modern Genetic Analysis: Integrating Genes and Genome*, W.H. Freeman and Company,2/e, 2002.

[3] Jie Chen, Huai Li, Kaihua Sun, and Bill Kim, "How Will Bioinformatics Impact Signal Processing Research?," *IEEE Signal Processing Magazine*, pp.16-26, Novermber 2003.

[4] Dimitris Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, Vol.16, no.12, pp.1073-1081, 2000.

[5] David S. Stoffer, "Spectral analysis for categorical time series: Scaling and the spectral envelope," *Biometrika*, Vol. 80, No. 3, pp.611-622, 1993.

[6] V.R. Chechetkin and A. Y. Turygin, "Size-dependence of three-periodicity and long-range correlations in DNA sequences," *Phys. Lett. A*, vol. 199, pp.75-80, 1995.

[7] J.W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, pp.5303-5318, 1982.

[8] S. Tiwari, S. Ramachandran, A. Bhattacharya., S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *CABIOS*, vol. 113, pp.263-270, 1997.

[9] D. Anastassiou, "Genomic Signal Processing," *IEEE Signal Processing Magazine*, July 2001

[10] Maniewski R., Lexandowski P., Liebert A., Mroczka T., and Steinbach K., "Spectral Analysis of High Resolution ECG In Study On Late Potentials," *Proceedings RC IEEE-EMBS & 14th BMESI*, 1995.

[11] Gen Bank Available: http://www.ncbi.nlm.nih.gov

[12] Susanna K. P. Lau, Patrick C. Y Woo, Mo-yin Mok, Jade L L Teng, Victoria K. P. Tam, Katie K. H. Chan, and Kwok-yung Yuen, "Characterization of Haemophilus segnis, an Important Cause of Bacteremia, by 16S rRNA Gene Sequencing," *Journal of Clinical Microbiology*, pp.877-880, Feb. 2004.