Low Complexity Frequency Monitoring Filter for Fast Exon Prediction Sequence Analysis

Daniel Massicotte[1], Marwan A. Jaber[1], Marie-Ange Massicotte[2], Philippe Massicotte[1]

[1]Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada, G9A 5H7
Electrical and Computer Engineering Department, Laboratory of Signal and System Integrations
{daniel.massicotte, marwan.jaber, philippe.massicotte2}@uqtr.ca

[2]Université Laval, Département de biochimie, microbiologie et bio-informatique, Québec, Canada
Marie-Ange.Massicotte.1@ulaval.ca

***Abstract* – Over the last few years, the application of Digital Signal Processing (DSP) techniques for genomic sequence analysis has received great interest. Indeed, among its applications in genomic analysis, it has been demonstrated that DSP can be used to detect protein coding regions (exons) among non-coding regions in a DNA sequence. The period-3 behavior exhibited by exons is one of its features that has been exploited in several developed algorithms for exon prediction. Identification of this periodicity in genomic sequences can be done by using different methods such as the well-known Fast Fourier Transform (FFT) and the Goertzel algorithm for complexity reduction in which the reduction of computational time is a great challenge in genomic analysis. Therefore, this paper presents a novel one frequency analysis by using half of the arithmetic complexity of the Goertzel algorithm for gene prediction. Compared to the Intel®'s FFT (MKL) optimized function, the Goertzel's (IPP) and the dedicated Goertzel compiled function with ICC on Xeon CPU (24 cores), the proposed method conserves the same accuracy provided by the referenced methods which will manifest a speedup of 3000, 10 and 2 compared to MKL FFT, IPP Goertzel and the dedicated Goertzel with ICC, respectively.**

Keywords: DNA sequence, Fast Fourier transform, Goertzel algorithm, Period-3 behavior, One frequency filter.

1. Introduction

DNA (deoxyribonucleic acid), the carrier of all genetic information, consist of long strings of four subunits called nucleotides (adenine, thymine, guanine, and cysteine) [1]. The specific order of these subunits forms the human genes, which bears all the instructions necessary for organisms to develop, survive and reproduce. Indeed, through complex mechanisms, genes dictate, by their DNA sequence, the structure of proteins, molecules responsible for a multitude of molecular function in organisms. Identification and understanding the functions of genes is therefore indispensable for fundamental biology research as well as applied fields such as medicine. Unfortunately, to predict the regions of a DNA sequence that encode for a gene and ultimately a protein is a challenging task, especially in the case of eukaryotic studies. This complexity is caused by the discontinuous feature of eukaryotic genes where protein coding regions of the sequence, called exons, are divided into smaller segments by non-coding regions such as introns [2][3]. Thus, in order to accurately predict exons positions on a DNA sequence, specific features of protein coding regions need to be exploited.

In recent year, a widely used feature in exon prediction algorithm is the three-base periodicity pattern found in exons. This pattern is absent in non-coding regions which exhibit a rather random pattern. This difference between coding and non-coding regions has been confirmed through the Fourier analysis of DNA sequences, where coding regions display a prominent peak at the frequency $N/3$, while no such peak is observed in non-coding regions [4][5] where N is the length of the coding region. Thus, frequency analysis can be used to distinguish exons from non-coding regions in a DNA sequence, which can subsequently lead to gene identification.

Over the years, several computational tools have been developed for exon prediction based on frequency analysis. Among those, we found methods based on the Fast Fourier Transform [1],[2],[6] or the Goertzel algorithm [19] to compute the period-3 pattern for exon prediction sequence analysis. When only one frequency of the sequence N needs to be computed, the Goertzel's algorithm represents the lower complexity [19]. Methods largely used in many applications. To our knowledge, the authors of [7] are the first to propose Goertzel for DNA analysis. Many other works based on Goertzel's algorithm have been proposed [9][10][11]. Someone concerned the implementation on CPU multicore and GPU [10], others have implemented it on a field programmable gate array (FPGA) [9][11] to accelerate the computation.

With the great length and complexity of the sequences that need to be analyzed (e.g. the human genome is 3 billion DNA base pairs). As a result, with the increasing number of the sequenced genomes being and the rising interest in conducting comparative genomic analysis that will involve comparing thousands of sequences, it is extremely relevant to develop and optimize exon prediction tools with highly fast processing speed. This is especially important in case of prediction methods based on frequency analysis, since they are generally used in combination with other prediction tools, and thus their computational time need to be minimal. In this regard, the aim of this study was to develop an algorithm for exon prediction that is less time-consuming while conserving the reliability in prediction's accuracy offered by the Goertzel algorithm.

To reduce the processing time, we will be proposing to a novel algorithm, JM-Filter[1], that is dedicated to compute one frequency with less complexity which is offered by the Goertzel algorithm. The Goertzel's algorithm needs N iterations for each computed frequency in the sequence. In this paper, we are proposing one frequency filter that will reduces by a factor of 2 the number of iterations yielding to a computational time reduction while maintaining the same accuracy as the Goertzel algorithm.

This paper is organized as follows: Section 2 will briefly detail Genomic Signal Processing (GSP), meanwhile Section 3 elaborates the proposed first and second order JM-Filter, Section 4 draws on the performance results in terms of exon detection and computational time, and finally a conclusion in Section 5.

2. Genomic Signal Processing

As previously mentioned, a DNA sequence is made of four nucleotides each of which is symbolized by a unique alphabetical character; adenine (A), thymine (T), guanine (G) and cytosine (C). Thus, from a mathematical point of view, a DNA sequence can be regarded as a long string composed of these four characters (A, T, G, and C). To performs a frequency analysis of a DNA sequence, the first step is to transform its alphabetical representation into a binary representation, so it can be treated on a digital signal processor (DSP).

In this paper, we adopt the following representation: if the DNA sequence ATCG contains the nucleotide *A,* therefore the vector A is coded as

$$A=\left[x_A(n)=1 \quad x_T(n)=0 \quad x_C(n)=0 \quad x_G(n)=0\right] \tag{1}$$

else if the DNA sequence ATCG contains the nucleotide T therefore, the vector T is coded as

$$T=\left[x_A(n)=0 \quad x_T(n)=1 \quad x_C(n)=0 \quad x_G(n)=0\right] \tag{2}$$

On the other hand, if the DNA sequence ATCG contains the nucleotide C therefore, the vector C is coded as

$$C=\left[x_A(n)=0 \quad x_T(n)=0 \quad x_C(n)=1 \quad x_G(n)=0\right] \tag{3}$$

meanwhile, if the DNA sequence ATCG contains the nucleotide G therefore, the vector G is coded as

$$G=\left[x_A(n)=0 \quad x_T(n)=0 \quad x_C(n)=0 \quad x_G(n)=1\right] \tag{4}$$

A summarized binary representation of a DNA sequence is presented in Table 1.

Table 1 Converting the DNA sequence into binary sequence

Sequence	A	C	T	C	A	G	G	T
$x_A(n)$	1	0	0	0	1	0	0	0
$x_T(n)$	0	0	1	0	0	0	0	1
$x_C(n)$	0	1	0	1	0	0	0	0
$x_G(n)$	0	0	0	0	0	1	1	0

By adopting $X_A(k)$, $X_T(k)$, $X_C(k)$ and $X_G(k)$, as the DFT (Discrete Fourier Transform) of the DNA's binary sequences $x_A(n)$, $x_T(n)$, $x_C(n)$ and $x_G(n)$; the DNA character string at a frequency k is obtained by the power spectral P(k) expressed as [8][12]:

$$P(k)=\left|X_A(k)\right|^2+\left|X_T(k)\right|^2+\left|X_C(k)\right|^2+\left|X_G(k)\right|^2 \tag{5}$$

where $k = 0, 1\ldots, N-1$.

As previously stated, a coding region of a gene (exon) exhibits a period-3 pattern that is translated into large peaks in the spectral domain occurring at $k = N/3$ of the DFT coefficients. To speed-up the computation of the exon coding region, Short Time Discrete Fourier Transform (STDFT) is applied with a sliding window according to:

$$X_\alpha(k)=\sum_{n=0}^{N-1} w(n)\,x_\alpha(n)\,e^{\frac{-2\pi kn}{N}} \tag{6}$$

where α = A, C, G, T, and $w(n)$ is a rectangular window given by:

$$w(n)=\begin{cases}1 & for\ 0<n<N-1\\ 0 & elsewhere\end{cases} \tag{7}$$

Different windows, could be applied, but we will stick with the rectangular window.

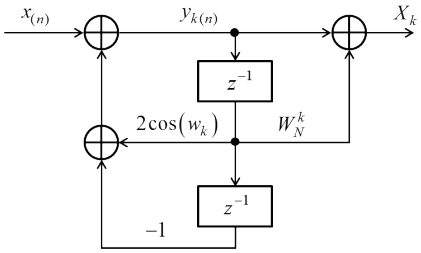

Fig. 1 The second order Goertzel algorithm.

The second order Goertzel algorithm is shown in Fig 1. This algorithm compute one specific frequency from the input sequence $x_{(n)}$ using only N recursions. The arithmetic complexity is reduced to N compared to the FFT using $N\log N$.

The main contribution of this paper is to show the reduction of complexity and computational time compared to the Goertzel algorithm (Fig. 1) of exon prediction based on the novel JM-Filter to detect specific frequencies in a monitored signal.

3. The proposed First and Second Order JM-Filter

This section briefly describes the radix 2 JM-Filter algorithm. For more information, all details are available in [13], where the radices 4 and 8 and the accuracy in fixed-point are analyzed. The one iteration FFT algorithm expressed as [14][15][16]:

$$X_{(qV+v)}=\sum_{p=0}^{V-1} W_N^{[\![p(qV+v)]\!]_N}\left(x_{(p)}+x_{(V+p)}W_N^{[\![vV]\!]_N}\right), \tag{8}$$

[1] "Filter Configured to Detect Specific Frequencies of a Monitored Signal ", US Patent Application No 62/677,587, 2018.

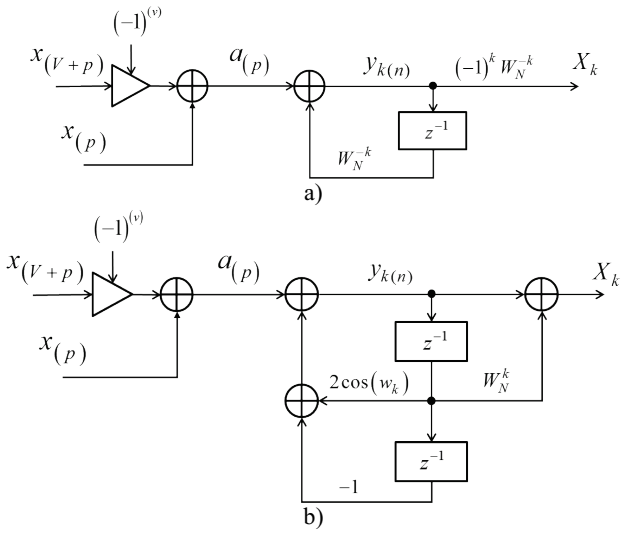

Fig. 2 The proposed JM-Filter radix-2 first (a) and second order (b).

where $[\![x]\!]_N$ represents the operation x modulo N, $v = 0,1,....,V-1$, $q = 0,1$, and $V = N/2$.

To compute a specific frequency $X_{(k)}$ for a given k, the values of q and v must be known in advance.

For such types of FFTs, the k domain is subdivided into 2 equal sub-domain of size $N/2$ as presented in [15] and in order to compute a specific frequency $X_{(k)}$ for a given k the values of q and v should be known in

$$\begin{cases} 0 \le k < V & q = 0 \text{ and } v = k \\ V \le k < N & q = 1 \text{ and } v = k - V \end{cases} \tag{9}$$

We can define the second part of the Eq. (8) as follow:

$$a_{(p)} = \left(x_{(p)} + x_{(V+p)} W_N^{[\![vV]\!]_N}\right) = \left(x_{(p)} + x_{(V+p)} e^{-j\pi mv}\right), \tag{10}$$

By examining Eq. (10), further reductions in terms of complexity could be achieved for the radix-2 case, since

$$e^{-j\pi mv} = (-1)^{mv}. \tag{11}$$

Therefore, based on Eq. (11), we can re-write Eq. (10) as

$$a_{(p)} = x_{(p)} + (-1)^v x_{(V+p)}, \tag{12}$$

the radix-2 JM-Filter first order filter would be

$$y_{k(p)} = W_N^{-k} y_{k(p-1)} + x_{(p)} + (-1)^v x_{(V+p)} \tag{13}$$

and the k^{th} computed frequency is given by (as shown in Fig. 2a)

$$X_k = (-1)^k W_N^{-k} y_{k(V-1)}. \tag{14}$$

The radix-2 second-order JM-Filter will be (Fig. 2b)

$$y_{k(p)} = 2\cos(2\pi k / N) y_{k(p-1)} - y_{k(p-2)} + a_{(p)}, \tag{15}$$

where $y_{k(-2)} = y_{k(-1)} = 0$, and from which the k^{th} computed frequency is:

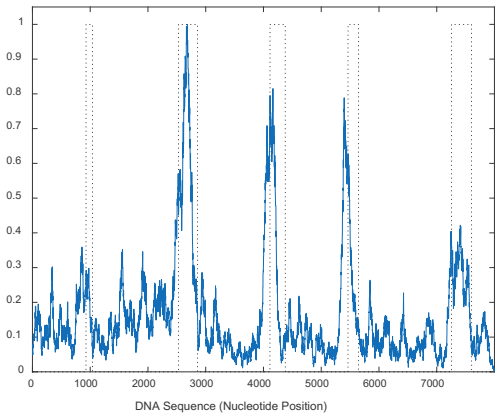

Fig. 3 Detection of period-3 behavior using FFT, Goertzel and the proposed JM-Filter. All results are superposed to shown the same accuracy. The dashed lines describe the ideal gene positions.

$$X_k = (-1)^k \left(0.5\cos(2\pi k / N) y_{k(V-1)} + \sin(2\pi k / N) - y_{k(V-2)}\right) \tag{16}$$

Based on Fig. 2 which shows the first and second order, we can clearly notice that the computed number of iterations in Eqs. (12) and (15), depends on p=0,1,…,V and (V=N/2). Compared to Goertzel algorithm, the number of iterations is reduced by a factor of 2. The summary of the computational complexity for the input sequence of size N is shown in Table 2 for Goertzel algorithm and the proposed JM-Filter. The evaluation is done on DNA sequences and the computational time is done using C language on Xeon processor.

Table 2 Computational complexity in terms of real arithmetic operations for the first and second order for an input sequence signal of length N.

Methods	First Order		Second Order	
	Mult	Add	Mult	Add
Goertzel [19]	4N	4N	2N+2	4N-2
JM-Filter	2N	3N	N+2	3N -2

Considering our interest to detect only one frequency in the window, $k = N/3$, and the equation (16) becomes

$$y_{k(p)} = y_{k(p-1)} - y_{k(p-2)} + a_{(p)}, \tag{17}$$

From Eq. (17) we can conclude that the summation of $x_{(p)} - x_{(V+p)}$ can be computed only one time outside the recursion of the complete sequence M. The result $a_{(p)}$ becomes the input of the JM-Filter with a length of $N/2$. The square computation of the k^{th} output frequency detected according to equation (5), can be executed using only real value for the second order Filter which will simplify the memory's access based on equation (16).

4. Performance Results

Our performance study compares our results to the methods cited in [11][10]. In our simulations we used the same genomic sequence cited in [17]. The sequence of 8000 nucleotides (M=8000) and encoding for the F56F11.4 gene of

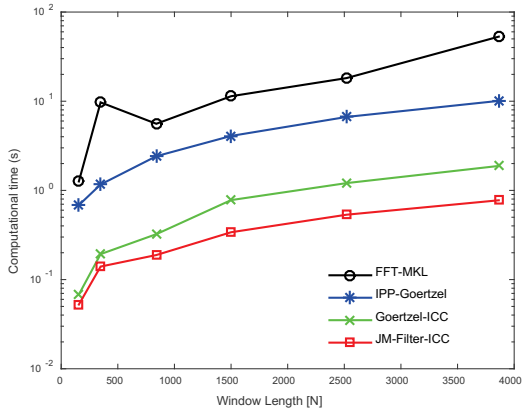

Fig. 4 Computational time for proposed and Goertzel methods, both for second order algorithm, and the FFT for different window sizes *N*.

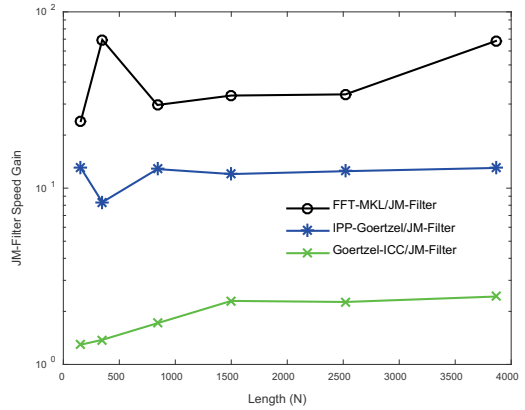

Fig. 5 Computational time ratio between our proposed method and reference methods (MKL FFT, IPP Goertzel, and Goertzel with ICC) for different window sizes *N*.

C.elegans was directly extracted from the Genbank database maintained by the National Biotechnology Information Center (NCBI) [18]. Since this targeted gene is composed of five exons [1], five peaks are expected to be detected following the analysis of the period-3 behavior using the FFT technique, the Goertzel's algorithm and our proposed JM-Filter. The detection of the 3-period behavior by implementing those three methods (FFT, Goertzel and JM-Filter). The estimation with a normalized magnitude is illustrated in Fig. 3 in which all these methods have predicted the same period-3 behavior for the F56F11.4 gene. The prediction is based on a window size of 348 for all methods.

The window size affects the accuracy of exon prediction sequence. To predict a large exon sequence, we need to use a large window size, conversely, for short exon sequence prediction, we need to use a short window size. The time consuming is directly affected by the window size *N* and the DNA sequence *M* analysis.

The window size affects the accuracy of the exon prediction sequence. To predict a large exon sequence, we need to use a larger window size, meanwhile for shorter exon sequence prediction, we need to use a shorter window size. The execution time is directly affected by the window size *N* and the DNA sequence *M* analysis.

The sketched results in Figs (4 and 5), are executed on two Intel® processors Xeon® CPU E5-2620 v3 with 12 cores (total of 24 cores) at 2.40 GHz and 258 GB of RAM memory. Goertzel and JM-Filter algorithms are coded in C language and compiled on the Intel® C 64' compiler for applications running on Intel® 64 (ICC) version 19.0.4.243 build 20190416. To evaluate the computational time, the used genomic sequence was the chromosome III of *C.elegans* of length *M*=13 783 801 (Genbank accession number NC_003281) [18]. The window sizes 348, 1500, 2520 and 3 864 were applied to test the computational time effects. The optimized MKL coded FFT function and the IPP Goertzel algorithm coded for Intel® are used to compare the performances.

Fig. 4 shows the execution time for different window sizes with MKL-FFT, IPP-Goertzel second order, Goertzel-ICC second order, and JM-Filter-ICC second order. A significant gain is observed compared to the FFT and the IPP-Goertzel. The IPP Goertzel corresponds to the Goertzel algorithm as a function developed and optimized by Intel®. The IPP Goertzel is a general function dedicated to Goertzel and assuming the N recursions using the coefficient $2\cos(w_k)$ with no simplification like equation (17). Fig. 5 reveals the computational time ratio between Goertzel and our proposed method for the first and second order. For a window size more than 1 500, the gains could be over of 2.2 in speed for the proposed JM-Filter second order method in comparison to dedicated ICC Goertzel algorithm. The proposed method presents a complexity reduction of $N/2$ recursions compared to Goertzel. Thus, the gain of computational time exceeded 2 is due to the saving in the memory's access.

Table 3 shows the computation accuracy for all methods in double precision on Xeon CPU using different window sizes *N*. The accuracy is computed based on the relative error based on the norm L_2.

Table 3 The computation accuracy in double precision for the proposed and the cited methods for different window sizes *N*.

Window Size	IPP Goertzel	Goertzel ICC	JM-Filter ICC
348	1,37E-02	1,61E-02	5,45E-04
1500	4,23E-03	4,57E-03	1,20E-02
2520	3,47E-03	3,64E-03	1,19E-02
3864	2,75E-03	2,85E-03	8,83E-03

5. Conclusion

This paper has finally introduced a filter suitable for genomic signal processing that can reduce the computational complexity compared to Goertzel's algorithm. The proposed filter reduces the number of iterations by a factor of 2 and computational speed-up by a factor of 2.2 for the second order filter. Our future work will be mainly concentrated on the

implementation of our proposed filter on NVIDIA's Jetson GPU card and FPGA devices in which the computation of a specific frequency will occur.

Acknowledgements

This work has been funded by the Natural Sciences and Engineering Research Council of Canada grant, the Canada Foundation for Innovation, the CMC Microsystems and the Research Chair in Signals and Intelligence of High Performance Systems. The authors would also like to thank Loïc Bachelot for his help in the implementation work.

References

[1] A. Dimitris, "Genomic signal processing", IEEE Signal Processing Magazine, 2001, pp. 8-20.

[2] C. Yin, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence", Journal of Theoretical Biology, April 2007, pp. 687-694.

[3] J.D. Watson et al., Molecular Biology of the Gene, Pearson, 2014, 872 p.

[4] B.D. Silverman and R. Linsker "A measure of DNA periodicity," Journal of Theoretical Biology, vol. 118, February 1986, pp. 295-300.

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

[6] S. Marhon and S.C. Krener, "Gene prediction based on DNA spectral analysis: a literature review," Journal of Computational Biology, 2011.

[7] A.R. Fuentes et al., "Detection of coding regions in large DNA sequences using the short time Fourier transform with reduced computational load," Progress in Pattern Recognition, Image Analysis and Applications. Lecture Notes in Computer Science, Springer, 2006, pp. 902-906.

[8] M. Abo-Zahhad, S. M. Ahmed and S. A. Abd-Elrahman, "Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques," J. Information Technology and Computer Science, 2012, 22-36.

[9] H.T. Bui, "Pipelined FPGA design of the goertzel algorithm for exon prediction", IEEE International Symposium on Circuits and Systems, Seoul, South Korea, 2012, pp. 572-575.

[10] S.R. Rivard, J.G. Mailloux, R. Beguenane, H.T. Bui, "Design of high-performance parallelized gene predictors in MATLAB," BMC Research Notes, April 2012, pp. 1-10.

[11] M. Voyer et al, "Rapid prototyping of the Goertzel algorithm for hardware acceleration of exon prediction," IEEE International Symposium of Circuits and Systems, Rio de Janeiro, Brazil, 2011, pp. 85-88.

[12] G. Chen, X.-Ming Dou, and X.-Fang Zhu, "Extron prediction method based on improved period-3 feature strategy, Modern Physics Letters B, Vol. 31, No 19, 2017, pp. 1-6.

[13] M. Jaber and D. Massicotte, "The JM-Filter to Detect Specific Frequencies in Monitored Signal," submitted to IEEE Trans. On Signal Processing, June 2019.

[14] M. Jaber and D. Massicotte, "The Radix-*r* one stage FFT kernel computation," Int. Conf. Acoustic, Speech, and Signal Processing, April First, Las Vegas, Nevada USA, 2008, pp. 3585-3588.

[15] M. Jaber, D. Massicotte, "Fast method to detect specific frequencies in monitored signal", International Symposium on Communications, Control and Signal Processing, Cyprus, March 2010, pp. 1-5.

[16] M. Jaber and D. Massicotte, "A Novel approach for FFT data reordering," International Symposium on Circuits and Systems, Paris, May 2010, pp. 1615-1618.

[17] M. Akhtar, E. Ambikairajah, and J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," IEEE International Conference on Emerging Technologies, 2005, pp. 13-17.

[18] NCBI GenBank database, online access: http:Hlwww.ncbi.nlm.nih.gov/Genbank/

[19] G. Goertzel, "An algorithm for the evaluation of finite trigonometric series", American Mathematical Monthly, 1958, pp. 34-35.