Research, Society and Development, v. 9, n. 9, e883998173, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i9.8173

Identification of exonic regions in dna sequences an approach using cross-correlation and noise suppression by discrete cosine transform

Identificação de regiões exônicas em sequências de dna uma abordagem usando correlação cruzada e supressão de ruído por transformação de cosseno discreta

Identificación de regiones exónicas en secuencias de adn un enfoque utilizando correlación cruzada y supresión de ruido mediante transformación discreta de coseno

Received: 09/04/2020 | Reviewed: 09/06/2020 | Accept: 09/10/2020 | Published: 09/11/2020

Aratã Andrade Saraiva
ORCID: https://orcid.org/0000-0002-3960-697X
Institute of Mathematics and Computer Sciences, University of São Paulo, Brazil
E-mail: aratasaraiva@gmail.com
Felipe Castro
ORCID: https://orcid.org/0000-0002-7751-9455
Department of Computing, University of Piauí State, Brazil
E-mail: fenris@prp.uespi.br
Marcos Soares de Oliveira
ORCID: https://orcid.org/0000-0003-2389-3334
Department of Computing and Mathematics, University of São Paulo, Brazil
E-mail: marcossoares10@usp.br
José Vigno Moura Sousa
ORCID: https://orcid.org/0000-0002-5164-360X
Department of Computing, University of Piauí State, Brazil
E-mail: fenris@prp.uespi.br

Abstract

To identify the exonic regions in the DNA sequence of Chromosome 23, filtering techniques are used. DCT is a technique with the ability to remove noise from signals as shown in [Saraiva et al., 2018], in addition, noise suppression with DCT is not enough in itself, so in this work a new method of identifying exonic regions using cross correlation with DCT together with an FFT-based bandpass filter to decrease signal noise and find exonic regions.

Keywords: DNA; Exonic region; Signal processing.

Resumo

Para identificar as regiões exônicas na sequencia de DNA do Cromossomo 23, técnicas de filtragem são utilizadas. O DCT é uma técnica com capacidade de retirar ruído de sinais como mostrado no [Saraiva et al., 2018], além disso, a supressão de ruído com DCT não é suficiente por si só, então neste trabalho um novo método de identificar as regiões exônicas usando correlação cruzada junto com DCT juntamente com um filtro de passa banda com base em FFT para diminuir o ruído do sinal e encontrar as regiões exônicas.

Palavras-chave: DNA; Região exônica; Processamento de sinais.

Resumen

Para identificar las regiones exónicas en la secuencia de ADN del cromosoma 23, se utilizan técnicas de filtrado. DCT es una técnica con la capacidad de eliminar ruido de señales como se muestra en [Saraiva et al., 2018], además, la supresión de ruido con DCT no es suficiente en sí misma, por lo que en este trabajo un nuevo método para identificar regiones exónicas usando correlación cruzada junto con DCT junto con un filtro de paso de banda basado en FFT para disminuir el ruido de la señal y encontrar regiones exónicas.

Palabras clave: ADN; Región exónica; Procesamiento de señales.

1. Introduction

The identification of protein coding regions (exons) in DNA sequences using signal processing techniques is an important component of bioinformatics and biological signal processing, this work presents a new identification method, in this case the cross correlation and the correlation coefficient was used to confirm the feasibility of the technique used. The availability of complete genome sequence of many eukaryotic organisms continues to contribute towards better understanding of their genome design and evolution. An average vertebrate gene consists of multiple small exons separated by introns that are 10 or 100 times longer. In order to understand the structure and evolution of eukaryotic genomes, it is important to know the general statistical characteristics of the exons and introns, furthermore the identification of the exonic regions assist in the process of analyzing the eukaryotic genome sequence (Avery et al., 1944; Morgan, 1911).

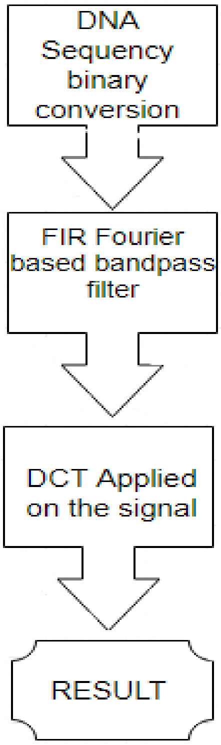
When the DNA sequence of a new eukaryotic organism is synthesized, the exonic (protein coding) regions must be distinguished from the introns. The protein coding regions of

DNA have been observed to exhibit a period-3 property due to the non-uniform codon usage in the translation of codons into amino acids (Fickett, 1982). The aim of this paper is to use this property to identify exonic regions (Hershey & Chase, 1952). A few codons take an interest more in protein union than others, offering ascend to reiterations of a particular sort of codon in the genome. For instance, the presence of an enormous number of GCA codons in the exonic areas gives more noteworthy reiteration of G, C and A nucleotides in the primary, second and third codon position, individually (Shetty, 2018). As such, the G, C and A nucleotides show period-3 property in the exonic areas (Akhtar et al., 2008). Quality discovering techniques dependent on hereditary attributes, for example, advertiser, CpG Island, start and stop codon and so on, will in general be of deficient exactness. The portrayal of coding and noncoding locales dependent on nucleotide insights inside codons is depicted by, who utilized a 12-image letter set to recognize the fringes among coding and noncoding districts (Bernaola-Galv´an et al., 2000). Afterward, Nicorici and Astola sectioned the DNA grouping into coding and non-coding areas utilizing recursive entropic division and stop-codon measurements (Datta and Asif, 2005).

2. Methodology

In this section will be explained all the materials and methods used to achieve this work's results, to facilitate the understanding this section will be divided in 3 subsections that will explain in details the transforms and the statistics chosen. The step by step process is as shown in the Figure 1.

Figure 1. Methodology block diagram.

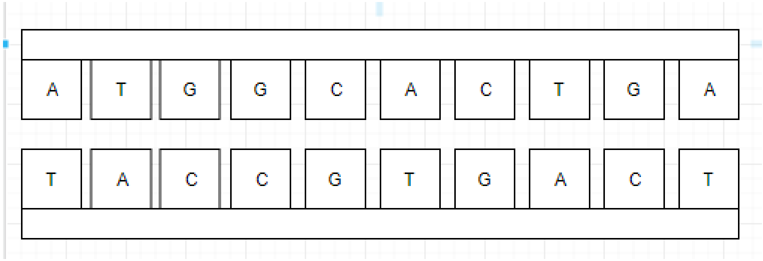

DNA Sequency binary conversion
FIR Fourier based bandpass filter
DCT Applied on the signal
RESULT

Source: Author.

2.1 DNA numeric conversion

To apply the technique to the DNA sequence in order to find nucleotide a region exhibiting a denoised signal, the DNA sequence is first mapped onto the numerical sequence, the DNA sequence is organized as shown in the Figure 2.

Figure 2. DNA sequence.



A	T	G	G	C	A	C	T	G	A
T	A	C	C	G	T	G	A	C	T

Source: Author.

The least complex transformation technique maps four mathematical successions IA[n], IT[n], IC[n] and IG[n] from DNA arrangements in paired organization. In this planning, the nearness or nonappearance of the particular nucleotides at the nth position is spoken to by '1' and '0', individually. For instance, given a segment of DNA succession ATCCGATATTC, the paired arrangement of the nucleotide A, signified IA[n], is [10000101000]. The paired groupings for the other three nucleotides T, C and G are discovered likewise After planning the DNA succession onto its parallel mathematical arrangement, the twofold grouping is gone through a Hamming window-based FIR channel of request 8 with focal recurrence set to 2/3. Absence of mutilations in FIR channels is one explanation behind their favored use over IIR channels in clinical applications.

Furthermore, the discrete cosine transform was applied on the signal to lower the noise on the data acquired, after that the signal become more understandable, and the exonic regions become more visible in the signal.

In the final step, the statistic is taken from the signal to find the statistic of how much accurate is the technique the statistic chosen was the cross-correlation and the correlation coefficient calculation of the resultant signal.

2.2 Fast Fourier Transform Based FIR Filter

Filters are signal conditioners. Each function by accepting an input signal, blocking pre specified frequency components, and passing the original signal minus those components to the output. For example, a typical phone line acts as a filter that limits frequencies to a range considerably smaller than the range of frequencies human beings can hear.

A digital filter takes a digital input, gives a digital output, and consists of digital components. In a typical digital filtering application, software running on a digital signal processor (DSP) reads input samples from an A/D converter, performs the mathematical manipulations dictated by theory for the required filter type, and outputs the result via a D/A converter. The FIR filter is designed using windowing, the method is to make an ideal filter in the frequency domain, and then translate it into the discrete time domain. However, this will give an infinite impulse response. To compensate for this, a window function is multiplied onto the ideal impulse response.

To make the ideal filter on the frequency domain we use the Fast Fourier Transform (FFT) and the hamming window as the principal tools. the FFT was defined like in the equation 1.

$$X(e^{j\omega}) = x(n)e^{-j\omega n}, \omega inradians \qquad (1)$$

In FIR filter design the order for the filter is denoted M and it determines the length of the window, corresponding to the discrete-time notation of h[n] as it shown in the equation 2.

$$h[n] = h_d[n] \cdot w[n] \tag{2}$$

As the approximated impulse response of the filter and with w [n] as the windowing function as is shown on the equation 3.

$$w[n] = \begin{cases} 1, & \text{for} 0 \leq n \leq M \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The product of (M.2) is in the frequency domain equal to the convolution as is shown below on the equation 4.

$$H(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_d(e^{j\omega}) \cdot W(e^{j\omega - \phi}) d\phi \tag{4}$$

The FIR frequency response H (ω) is a finite-degree polynomial in $e^{j\omega}$.

2.3 Discrete Cosine Transform

The discrete cosine transform (DCT) is very related to the Discrete Fourier Transform (DFT), it can often reconstruct a precise sequence of only a few DCT coefficients, this property is very useful for applications that require data reduction, precisely the purpose of this work, to explore the reduction of data use in electrocardiogram, [Nguyen et al., 2017].The DCT has four standard variants, for an x-signal of size N and with the kronecker δ, the transformations are defined by the equations 1, 2, 3 and 4 respectively.

$$y(k) = \sqrt{\frac{2}{N}} \sum_{N-1}^{N} x(n) \frac{1}{\sqrt{1+\ delta_{n1} + \delta_{nN}}} \frac{1}{\sqrt{1+\delta_{k1}\delta_{kN}}} cos\left(\frac{\pi}{N-1}(n-1)(k-1)\right) \tag{5}$$

$$y(k) = \sqrt{\frac{2}{N}} \sum_{n-1}^{N} x(n) \frac{1}{\sqrt{1 = \delta_{k1}}} cos\left(\frac{\pi}{2N}(2n-1)(k-1)\right) \tag{6}$$

$$y(k) = \sqrt{\frac{2}{N}} \sum_{n-1}^{N} x(n) \frac{1}{\sqrt{1+\delta_{n1}}} cos\left(\frac{\pi}{2N}(n-1)(2k-1)\right) \tag{7}$$

$$y(k) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} x(n) cos\left(\frac{\pi}{4N}(2n-1)(2k-1)\right) \qquad (8)$$

The series are indexed with n = 1 and k = 1 instead of the usual n = 0 and k = 0. On the equations, x is meaning the input array, y are the DCT itself and n is equal to the length of the transform, a positive integer scalar, with x and y being vectors (they can be matrices) (Nguyen et al., 2017).

In his work, Swarnkar using the standlet transform achieved better results compared to DCT and Wavelet transform, being able to illustrate well its results using data like SNR, also used in this work, CR and Price Related Differential (PRD), A.Swarnkar et al., 2017]. A DCT expresses a series of finitely many data points in terms of a sum of cosine functions oscillate at different frequencies. DCT has the applications of solving partial differential equations, Chebyshev approximation, audio compression, (Raj & Ray, 2017).

2.4 Correlation coefficient

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution. The correlation coefficient of two random variables is a measure of their linear dependence. If each variable has N scalar observations, then the Pearson correlation coefficient is defined as is shown in equation 9.

$$\rho(A, B) = \frac{1}{N-1} \sum_{i=1}^{n} \left(\frac{A_i - \mu_A}{\sigma_A}\right) \left(\frac{B_i - \mu_B}{\sigma_B}\right) \qquad (9)$$

Where µA and σA are the mean and standard deviation of A, respectively,and µA and σB are the mean and standard deviation of B. Alternatively, you can define the correlation coefficient in terms of the covariance of A and B as is shown in the equation 10.

$$\rho(A, B) = \frac{COV(A, B)}{\sigma A \sigma B} \qquad (10)$$

Several types of correlation coefficient exist, each with their own definitionand own range of usability and characteristics. They all assume values in therange from -1 to +1, where 1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

As tools of analysis, correlation coefficients present certain problems, including the propensity of some types to bedistorted by outliers and the possibility of incorrectly being used to infer a causal relationship between the variables. in this case the variables used was the signal and the exonic regions itself.

2.5 Cross-Correlation

In signal processing, cross-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other. This is also known as a sliding dot product or sliding inner-product. It is commonly used for searching a long signal for a shorter, known feature. It has applications in pattern recognition, single particle analysis, electron tomography, averaging, cryptanalysis, and neurophysiology. The cross-correlation is similar in nature to the convolution of two functions. In an autocorrelation, which is the cross-correlation of a signal with itself, there will always be a peak at a lag of zero, and its size will be the signal energy.

The result of a cross correlation can be interpreted as an estimate of the correlation between two random sequences or as the deterministic correlation between two deterministic signals.

The true cross-correlation sequence of two jointly stationary random processes, x_n and y_n, is given by the equation 10.

$$R_{xy}(m) = E\left\{x_{n+m}Y_n^*\right\} = E\left\{x_n Y_{n+m}^*\right\} \qquad (10)$$

Where −¡∞ n ¡∞ the asterisk denotes complex conjugation, and E is theexpected value operator, cross correlation can only estimate the sequence because, in practice, only a finite segment of one realization of the infinite-lengthrandom process is available.

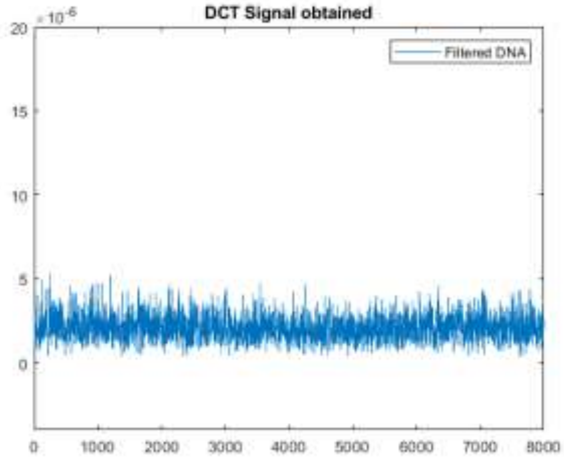
3. Results and Discussion

After the filters and statistics were applied the following results were obtained in each part of the work.

We will start the approach with the result of the identification of exonic regions, followed by the results of the statistics proving that this is an efficient technique, in the Figure 3.

is shown the signal obtained by the technique mentioned on this paper.

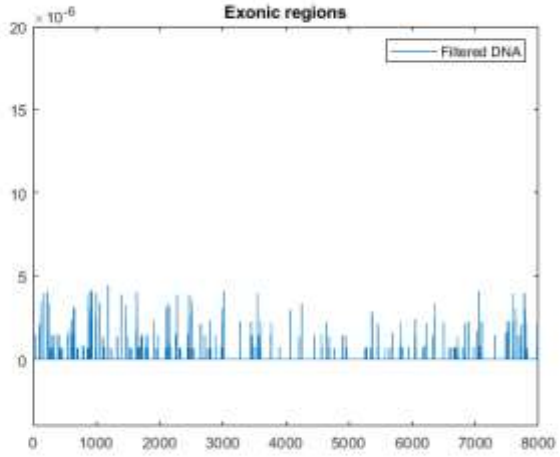
Figure 3. DCT reconstructed and FFT filtered DNA sequence.



Source: Author.

After that it is possible to highlight the exonic regions, the Figure 4 show the exonic regions present on the signal obtained as is shown on the figure the locations that are higher than zero are the exonic regions and the others are the introns.

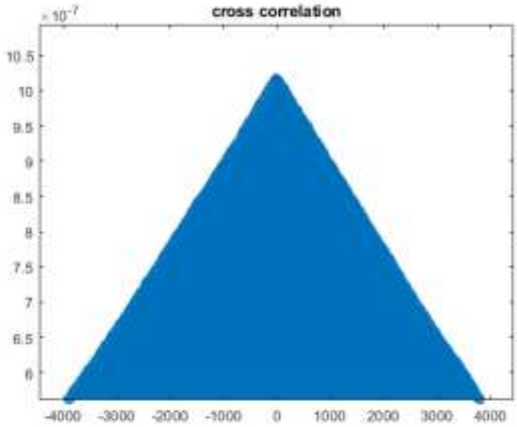
Figure 4. Highlited exonic regions.



Source: Author.

After that the cross-correlation are exemplified on the Figure 5, keeping in mind that the correlation coefficient also was used in the statistic, the result is shown on the Table 1.

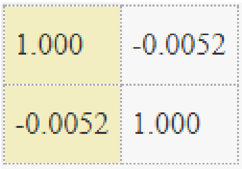
Figure 5. Signal and sequence cross correlation result.



Source: Author.

Observing the Figure 5, is possible to see that the cross correlation between theresult and the DNA sequence is possible find that the points are, grouped as follows, clearly resemblant a triangle as shows itself proving that the signal obtained has a high correlation to the first sequence, after that the Table 1 show the 1.000 result and a mean error of -0.0052 giving the accuracy of 99,48% which shows that the correlation to its original sequence is nearly optimal as mentioned on the statistic section.

Table 1. Correlation coefficient result and mean error



1.000	-0.0052
-0.0052	1.000

Source: Author.

To conclude this work, it was shown that DCT proved to be very effective for the reduction of noise in the signal obtained from the DNA sequence, even the correlation coefficient showed an excellent result, in terms of accuracy regarding the exons identified from signal obtained by DNA sequence which was shown in the Figure 3.

4. Final Considerations

Considering the results, the DCT used on this technique was proven itself truly reliable on this case to find the exonic results on the signal obtained by the geneticsequence after the FIR used to denoise it the exons was highlighted, to show the exon region on a easiest way to show the results with more emphasis.

Finally, for future works, the analysis of exonic regions is considered towork with using the DNA sequel of different chromosomes to find the difference on each exonic region and identify the peculiarity of itself.

References

Akhtar, M., Epps, J., & Ambikairajah, E. (2008). Signal processingin sequence analysis: advances in eukaryotic gene prediction.IEEE journal of selectedtopics in signal processing, 2(3),310–321.

A.Swarnkar, Kumar, R., Kumar, A., & Khanna, P. (2017). Per-formance of different threshold function for ecg compression using slantlet transform. insignal processing and integrated networks (spin). 37, 375–379.

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). Studies on thechemical nature of the substance inducing transformation of pneumococcal types: induc-tion of transformation by a desoxyribonucleic acid fraction isolated from pneumococcustype iii.Journal of experimental medicine, 79(2), 137–158.

Bernaola-Galv´an, P., Grosse, I., Carpena, P., Oliver, J. L.,Rom´an-Rold´an, R., & Stanley, H. E. (2000). Finding borders between coding andnoncoding dna regions by an entropic segmentation method.Physical Review Letters, 85(6):1342.

Datta, S. & Asif, A. (2005). A fast dft based gene predictionalgorithm for identification of protein coding regions. InProceedings. (ICASSP’05). IEEEInternational Conference on Acoustics, Speech, and Signal Processing, 2005., 5,653. IEEE.

Fickett, J. W. (1982). Recognition of protein coding regions in dna sequences.Nucleic acids research, 10(17), 5303–5318.

Hershey, A. D., & Chase, M. (1952). Independent functionsof viral protein and nucleic acid in growth of bacteriophage. The Journal of generalphysiology, 36(1),39–56.

Morgan, T. H. (1911). Chromosomes and associative inheritance.Science, 34 (880), 636–638.

Nguyen, B., Nguyen, D., Ma, W., and Tran, D. (2017). Investigatingthe possibility of applying eeg lossy compression to eeg-based user authentication. InNeural Networks (IJCNN), 2017 International Joint Conference on, pages 79–85. IEEE.

Raj, S. and Ray, K. C. (2017). Ecg signal analysis using dct-based dostand pso optimized svm. 66, 470–478.

Saraiva, A. A., Castro, F. M. J., Sousa, J. V. M., Valente, A., & Fonseca, F. (2018). Comparative study between the walsh hadamard transform anddiscrete cosine transform. In7th International Conference on Advanced Technologies.

Shetty, N. K. (2018). Inheritance of chromosomes, sex determination, and the human genome: A new paradigm.Gender and the Genome, 2(1):16–26.

Percentage of contribution of each author in the manuscript

Aratã Andrade Saraiva – 25%

Felipe Castro – 25%

Marcos Soares de Oliveira – 25%

José Vigno Moura Sousa – 25%