ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Method

Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach

Sitanshu Sekhar Sahu[1*] and Ganapati Panda[2]

[1]Department of Electronics and Communication Engineering, National Institute of Technology, Rourkela 769008, India;
[2]School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar 751013, India.

Genomics Proteomics Bioinformatics 2011 Apr; 9(1-2): 45-55 DOI: 10.1016/S1672-0229(11)60007-7

Received: Jul 19, 2010; Accepted: Oct 31, 2010

Abstract

Accurate identification of protein-coding regions (exons) in DNA sequences has been a challenging task in bioinformatics. Particularly the coding regions have a 3-base periodicity, which forms the basis of all exon identification methods. Many signal processing tools and techniques have been applied successfully for the identification task but still improvement in this direction is needed. In this paper, we have introduced a new promising model-independent time-frequency filtering technique based on S-transform for accurate identification of the coding regions. The S-transform is a powerful linear time-frequency representation useful for filtering in time-frequency domain. The potential of the proposed technique has been assessed through simulation study and the results obtained have been compared with the existing methods using standard datasets. The comparative study demonstrates that the proposed method outperforms its counterparts in identifying the coding regions.

Key words: protein-coding region, 3-base periodicity, time-frequency filtering, S-transform

Introduction

A major goal of genomic research is to understand the nature of the information and its role in determining the particular function encoded by the gene. A key step in achieving this goal is the identification of the gene locations and in deeper sense the protein-coding regions in the DNA strand. A DNA sequence is a long molecule that carries genetic information. It is composed of four types of different nucleotides, namely adenine (A), cytosine (C), guanine (G) and thymine (T). However, only some particular segments of the DNA molecule named as genes carry the coding information for protein synthesis. The complete genomes provide essential information for understanding gene functions and evolution. The determination of patterns in DNA and protein sequences is also useful for many important biological problems, such as identifying new genes, pathogenic islands and phylogenetic relationships among organisms. Hence accurate prediction of genes has always been a challenging task for computational biologists especially in eukaryote genomes.

The eukaryotic DNA is divided into genes and intergenic spaces. Genes are further divided into exons and introns. The exons carry the code for the production of proteins, hence they are called as protein-coding regions (*1-3*). It has been found that the bases in the protein-coding regions exhibit period-3 property due to the codon bias involved in the transla-

*Corresponding author.
E-mail: sitanshusekhar@gmail.com

tion process (*4-7*). This periodic behavior relates to the short term correlation in the coding regions. In addition, a long-range correlation (the so called 1/f spectrum) also presents in the genome sequence, which is considered as the background noise (*8, 9*). The presence of this noise makes the task of gene finding problem more complex. However, the 3-base periodicity (TBP) property has been used by many researchers as a good indicator of gene location. Rapid and accurate determination of the exon locations is important for genome sequence analysis. Computational approach is the fastest way to find exons in the genomic DNA sequences. Many techniques have been proposed and proved successfully in locating the protein-coding regions present inside the gene.

Several model-dependent methods like hidden Markov model (*10*), neural network (*11, 12*) and pattern recognition (*13*) have been successfully used to detect exons in genes. These models are supervised methods that are based on some prior information collected from the available databases. These methods are quite useful in the identification of coding regions, but not always. There may be a chance that the sequenced organism have coding regions that are not represented in the available databases. In addition, many model-independent methods have been proposed to identify the coding regions in DNA sequences. Basically these studies are based on the Fourier spectral content (*4, 14, 15*), spectral characteristics (*16*) and correlation of structure of DNA sequences (*8, 9*). Chakravarthy *et al* (*17*) and Akhatar (*18*) have proposed a parametric method of spectrum estimation based on autoregressive modeling. These methods require defining a prior analyzing window, within which the spectrum of DNA sequence is to be computed. As a result, it directly affects the efficiency and computational complexity of the predictor.

Hence there is a need for the development of alternative methods that reduce the window length dependency and should be efficient. Recently, Vaidyanathan and Yoon (*19, 20*) have proposed to use digital filters to identify the coding regions. In addition, Tuqan and Rushdi (*21*) have suggested a multirate DSP model for the same purpose. These model-independent methods do not require the prior window length and have shown to be effective in exon identification, but could not attain satisfactory accuracy level. Keeping these facts in mind, we have introduced a novel time-frequency filtering approach to this problem. This method is independent of the window length constraint and employs a time-band filter to extract the period-3 component in the DNA sequence and thereby identify the coding regions in it. It is also robust to the background noise present in DNA sequence. Case studies on genes from different organisms have demonstrated that this method can be an effective approach for exon prediction.

Materials and Methods

Data resources

In this work we focused on the analysis of eukaryotic DNA sequences that have been widely studied in the context of coding region identification. For demonstration purpose, we have used the DNA sequence of gene F56F11.4a (GenBank No. AF099922) on chromosome III of *Caenorhabditis elegans*. *C. elegans* is a free living nematode (roundworm), about 1 mm in length, which lives in temperate soil environment. It has been used as a benchmark problem for different gene detection techniques and known to have five distinct exons, relative to nucleotide position 7021 according to the NCBI database. The relative positions of the coding regions are 928-1039, 2528-2857, 4114-4377, 5465-5644 and 7265-7605. For the detailed analysis, we have also used the HMR195 benchmark dataset, which consists of 195 single gene sequences of human, mouse and rat (*22*).

Numerical mapping of DNA sequence

To apply suitable signal processing methods for the identification of protein-coding regions, the character string of the DNA sequence is converted to a suitable numerical sequence. This is achieved by assigning a numeral to each nucleotide that forms the DNA sequence. Hence, different techniques have been suggested to achieve this particular conversion. The aim of each coding method is to enhance the hidden information for further analysis. One of the most widely used mappings is the Voss mapping (*8*), where the character string of DNA is converted to four binary indicator sequences for each base (A, T, C and G). It assigns a numeral “1” when a particular symbol is found in the sequence, otherwise a “0”. Anastassiou (*14*) has proposed a complex number mapping by assigning a particular complex number to each base. Silverman and Linsker (*23*) have used a tetrahedron mapping, in which each nucleotide is assigned to one of the four corners of a regular tetrahedron. Chakravarthy *et al* (*17*) have proposed a real number mapping of the DNA sequence. Zhang *et al* (*24, 25*) presented a Z-curve mapping, which is a three-dimensional curve representation for the DNA sequence. Recently, Nair *et al* (*26, 27*) have used an EIIP indicator sequence to map the character string of DNA to numeric form. The EIIP is defined as the average energy of delocalized electrons of the nucleotide. Assigning the EIIP values to the nucleotides, a numerical sequence is obtained to represent the distribution of the free electrons’ energies along the DNA sequence. This has been successfully used to identify hot spots in proteins, for peptide design and also for identification of coding regions (*26, 28*). The EIIP sequence is a better choice for numerically representing DNA when compared to indicator sequences for the following reasons. First, it involves only a single sequence instead of four in the case of binary indicator sequences, thereby reduces the computational effort. Second, it is biologically more meaningful as it represents a physical property when compared to the indicator values, which represent just the presence or absence of a nucleotide. Hence in this paper, we have also used the EIIP representation method of numerical mapping of DNA sequence. The DNA sequence can be converted to the numerical sequence by replacing each nucleotide with the corresponding EIIP value. The EIIP values for the nucleotides are given in **Table 1**. For example, if x[n]=AATGCATCA, then using the values from Table 1, the corresponding EIIP numerical sequence is given as: x[n]=[0.1260 0.1260 0.1335 0.0806 0.1340 0.1260 0.1335 0.1340 0.1260].

Table 1 The EIIP values of nucleotides

Nucleotide	EIIP value
A	0.1260
T	0.1335
G	0.0806
C	0.1340

Spectral content measure method

In this frequency domain method, the discrete Fourier transform (DFT) of the EIIP indicator sequence is employed to exploit the TBP (*4, 14*). Let *U*[*k*] represents the DFT of the corresponding EIIP numerical sequence and is given by:

$$U[k] = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \qquad (1)$$

for *k* = 0, 1, …, *N*−1. Then the spectral content at k^{th} instant is:

$$S[k] = |U[k]|^2 \qquad (2)$$

S[*k*] acts as a preliminary indicator of a coding region giving a peak at the *N*/3 frequency. This procedure is used to detect the probable coding regions in the DNA sequence. Hence the coding regions are identified by evaluating *S*[*N*/3] over a window of *N* samples, then sliding the window by one or more samples and recalculating *S*[*N*/3]. This process is carried out over the entire DNA sequence. The peaks in the spectra obtained by the sliding window DFT correspond to the protein-coding regions. It is necessary that the window length *N* be sufficiently large (typical sizes are a few hundred to a few thousand), so that the periodicity effect dominates the background noise spectrum. This approach increases the computational complexity as it computes the spectrum within a window and is also constrained by the frequency resolution and spectral leakage effects of the windowed data record.

Digital filtering method

The Fourier-based spectral estimation method of exon identification can be viewed as a digital filtering perspective (*19*). The period-3 behavior of the coding regions is extracted by filtering the DNA sequence through a band pass filter *H*(*z*) with pass band centered at frequency 2π/3. The EIIP indicator sequence *x*(*n*) of the DNA sequence is passed through the filter *H*(*z*) to obtain the output sequence *y*(*n*), which contains the period-3 frequency. In the coding regions as it is expected to have period-3 component, a high energy particularly in these locations is produced. Thus to enhance this feature, the power of the filtered sequence is computed as:

$$Y(n) = [y(n)]^2 \tag{3}$$

Hence the plot of $Y(n)$ against "n" produces peaks in the coding regions and no peak in non-coding regions. The design and implementation of $H(z)$ as an anti-notch filter and its modifications have been discussed in many papers (*15, 20*). An overview of the implementation is presented as follows:

Consider a second-order all-pass filter

$$A(z) = \frac{R^2 - 2R\cos\theta z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \tag{4}$$

and a filter bank with two filters $G(z)$ and $H(z)$ defined as:

$$\begin{bmatrix} G(z) \\ H(z) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ A(z) \end{bmatrix} \tag{5}$$

Then $G(z)$ is defined as:

$$G(z) = k\left[\frac{1 - 2\cos\omega_0 z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}}\right] \tag{6}$$

where $G(z)$ is a notch filter with a zero at frequency w_0, when the radius R is less than and close to unity. Also the $H(z)$ and $G(z)$ are power complementary. Hence $H(z)$ can be a good anti-notch filter defined as:

$$H(z) = \frac{1}{2}\left[\frac{(1 - R^2)(1 - z^{-2})}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}}\right] \tag{7}$$

The DNA sequence can be viewed as a non-stationary signal, where the spectral components change along the sequence. Also it contains the background noise that comes due to the long-range correlations among the bases on the DNA stretch. Under such situation, the conventional Fourier domain filtering methods cannot extract properly the occurrence of period-3 component in the DNA sequence. Hence the joint time-frequency analysis (TFA) is needed for analyzing such spectral content in the sequence.

The proposed TFA method

TFA is of great interest when the signal models are unavailable. In such cases, the time or the frequency domain descriptions of a signal alone cannot provide comprehensive information for feature extraction and classification. Therefore, the time-frequency representation (TFR) (*29*) has evolved as a powerful technique to visualize signals in both the time and frequency domains simultaneously. Several techniques have been proposed for this purpose. Among them the short-time Fourier transform (STFT) (*30*) and the continuous wavelet transform (CWT) (*31*) are the most well known and widely used techniques. The major drawback in them is that STFT has a fixed resolution and CWT has a progressive resolution, but does not contain phase information (*31*). Recently in DSP literature, one efficient TFR tool, the S-transform (*32, 33*), has been proposed to possess superior time-frequency resolution as well as frequency detection capability.

S-transform: a TFR

S-transform is a TFA technique proposed by Stockwell *et al* (*32*), combining both properties of STFT and CWT. It provides frequency-dependent resolution while maintaining a direct relationship with the Fourier spectrum. The S-transform of a signal $x(t)$ is defined as:

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t) w(\tau - t, f) e^{-j2\pi ft} dt \tag{8}$$

where the window function is a scalable Gaussian window:

$$w(t, \sigma) = \frac{1}{\sigma(f)\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2(f)}} \tag{9}$$

and the width of the window varies inversely with frequency as:

$$\sigma(f) = \frac{1}{|f|} \tag{10}$$

Combining Formula (9) and (10) gives

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t)\left\{\frac{|f|}{\sqrt{2\pi}} e^{-\frac{(\tau - t)^2 f^2}{2}} e^{-j2\pi ft}\right\} dt \tag{11}$$

The advantage of S-transform over STFT is that the window width σ is a function of *f* rather than a fixed one as in STFT and thereby provides multi-resolution analysis. In contrast to wavelet analysis, the S-transform wavelet has a slowly varying envelope (the Gaussian window), which localizes the time and an oscillatory exponential kernel that selects the frequency being localized. It is the time localizing Gaussian that is translated while keeping the oscillatory exponential kernel stationary, which is different from the wavelet kernel. As the oscillatory exponential kernel is not translating, it localizes the real and the imaginary components of the spectrum independently, localizing the phase as well as amplitude spectrum. Thus it retains absolute phase of the signal, which is not provided by wavelet transform. Thus it better localizes the spectral content in the signal.

S-transform-based filtering approach

The standard Fourier-domain filtering techniques are constrained to stationary pass bands and reject bands that are fixed for the entire duration of the signal. These methods may be adequate for the stationary signals where the signal component of the data is time-independent and the noise is also time-independent. However, many signals are non-stationary in nature where the frequency response of the signal varies in time, or time-dependent noise exists. Hence there is a need for developing filters with time-varying pass bands and reject bands (*34, 35*). One of the most practical solutions to this problem is the joint time-frequency filter. In time-frequency filtering, the time frequency spectrum of a signal is first estimated, and portions that are part of the noise are removed using band-limited filters in those regions having the corresponding signal. In Formula (8), the S-transform window satisfies the condition:

$$\int_{-\infty}^{\infty} w(t,f)dt = 1 \qquad (12)$$

Therefore averaging the $S(\tau, f)$ over all values of t yields $X(f)$, the Fourier transform of $x(t)$:

$$\int_{-\infty}^{\infty} S(\tau,f)d\tau = X(f) \qquad (13)$$

The original signal can be recovered by using the inverse Fourier transform of $X(f)$:

$$x(t) = \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} S(\tau,f)d\tau\right\}e^{j2\pi ft}df \qquad (14)$$

Thus it provides a direct link between S-transform and Fourier transform. Due to the invertibility property of S-transform, it can be suitably used for time-frequency filtering. Let the signal $x(t)$ be a sum of main signal component $d(t)$ and noise component $n(t)$:

$$x(t) = d(t) + n(t) \qquad (15)$$

Due to the linearity property of S-transform, it is written as:

$$S(\tau,f) = D(\tau,f) + N(\tau,f) \qquad (16)$$

where D and N are the S-transform of the main signal and the noise, respectively. Therefore the filtering function $A(\tau,f)$ is to be such that:

$$D(\tau,f) = A(\tau,f)\cdot S(\tau,f) \qquad (17)$$

Using the inversion Formula (14), the denoised signal $\bar{x}(t)$ is recovered as:

$$\bar{x}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} D(\tau,f)e^{j2\pi ft}dtdf = \int_{-\infty}^{\infty} \bar{X}(f)e^{j2\pi ft}df \qquad (18)$$

Hence multiplying $S(\tau, f)$ with the filtering function $A(\tau,f)$ gives the S-transform of the denoised signal. In the present case, the period-3 signal in the genomic sequence is considered as the signal of interest and the rest is treated as noise. Hence the time-frequency filtering technique is used as a potential candidate to extract the protein-coding regions in the DNA segment.

Identification of protein-coding regions in DNA sequence using S-transform-based filtering approach

Nucleotides are assigned by the corresponding EIIP value as given in Table 1, which provides a numerical form of the DNA sequence. Then the spectrum of the DNA sequence under consideration is computed to observe the distribution of the energy of the frequency components throughout the sequence. In the spectrogram the bright regions correspond to the high energy areas relevant to the frequencies present. It has been observed that the period-3 frequency is the dominant frequency present in the coding regions of the DNA sequence along with some insignificant frequencies. For illustration purpose, the spectral distribution obtained by the proposed method for the gene F56F11.4a of *C. elegans* chromosome III is shown in **Figure 1**. The distinct energy-concentrated areas (bright areas) corresponding to the period-3 frequency

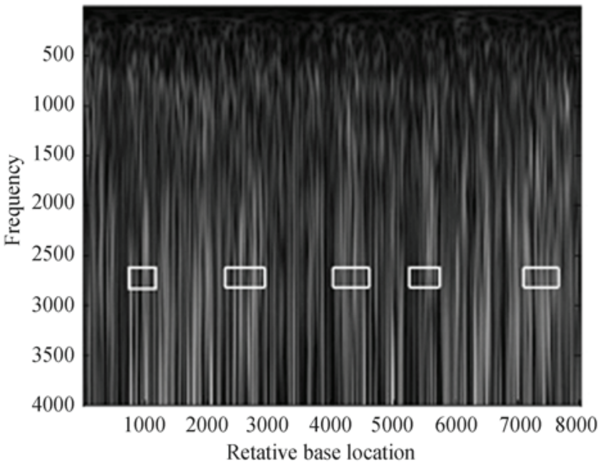

Figure 1 Spectrogram of the DNA sequence of F56F11.4a. The high spectrum values corresponding to period-3 frequency relevant to the coding regions are indicated by the rectangular boxes in the time-frequency plane.

in the time-frequency plane are indicated by the rectangular boxes. Then a specific band-limited time-frequency filter (mask) is designed to separate the period-3 frequency of interest.

The complete step-by-step procedure of the proposed S-transform-based filtering for identification of the coding regions is outlined in sequence:

1. Convert the DNA sequence of interest into a numerical sequence using the EIIP values (Table 1).
2. Compute the spectrum of the DNA sequence using the S-transform as defined in Formula (11).
3. Design the band-limited filter (mask) in time-frequency domain, which selects the period-3 frequency and activates during the specific regions in the time-frequency plane.
4. Filter the DNA numerical sequence of interest by using the time-frequency filter.

The peaks in the energy of the filtered output signal identify the locations of the protein-coding regions. If the output signal is denoted as $y(n)$, then its energy is given as:

$$E(n) = |y(n)|^2 \tag{19}$$

This energy is referred to as the energy sequence corresponding to TBP of the DNA sequence. Then the coding regions are predicted by the threshold of the energy sequence. The whole process of the proposed filtering approach for hot spot identification is presented in a flow graph in **Figure 2**.

Evaluation

To demonstrate the performance of the proposed method, the DNA sequence of the gene F56F11.4a of *C. elegans* chromosome III is analyzed. In this paper the existing model-independent methods, such as conventional sliding window DFT and the IIR anti-notch filter, are also simulated, and the results obtained are compared with those obtained by the proposed method. The simulation results of this particular gene are presented in **Figure 3** for comparison purpose.

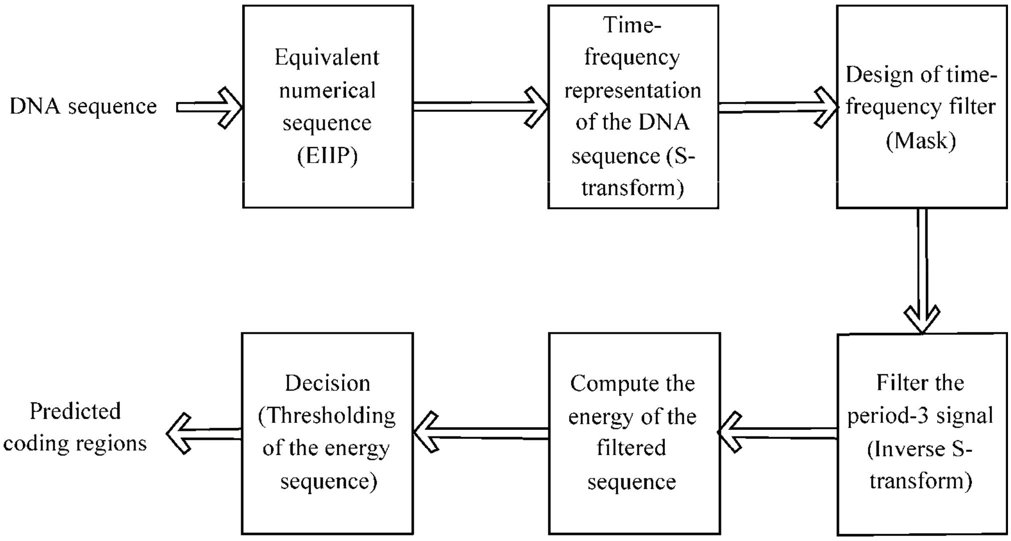
DNA sequence
Equivalent numerical sequence (EIIP)
Time-frequency representation of the DNA sequence (S-transform)
Design of time-frequency filter (Mask)
Filter the period-3 signal (Inverse S-transform)
Compute the energy of the filtered sequence
Decision (Thresholding of the energy sequence)
Predicted coding regions

Figure 2 The flow graph of S-transform-based filtering approach for protein-coding region identification.

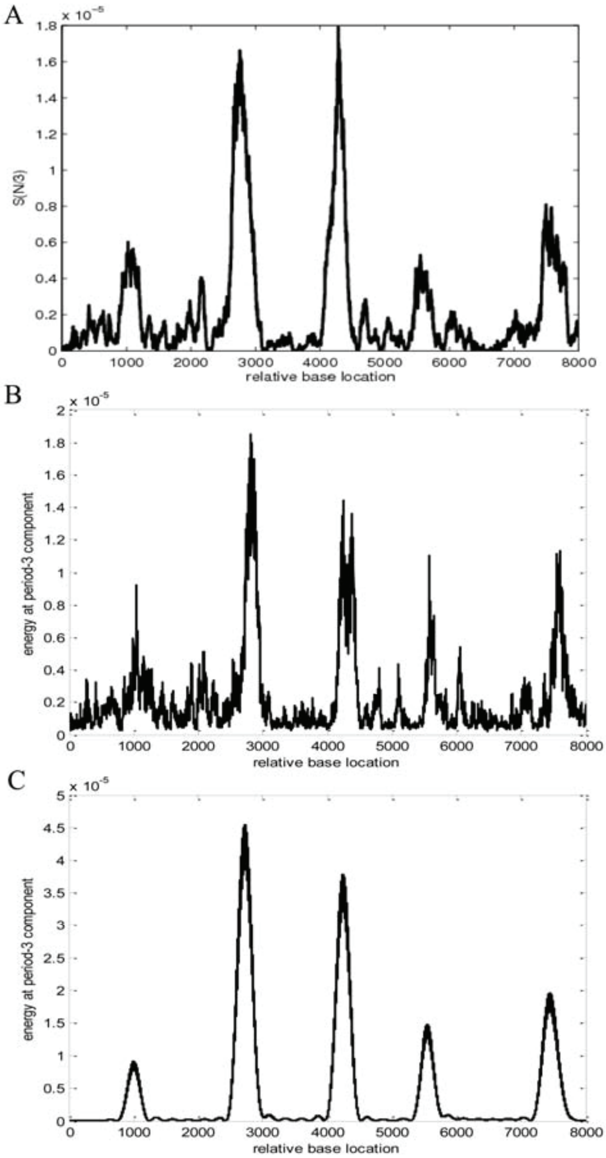

Figure 3 Comparison of the power spectra of the gene F56F11.4a obtained by DFT (**A**), anti-notch filter (**B**) and S-transform filter (**C**).

In the DFT spectrum analysis, a rectangular window of length 351 bp and step size of 1 is used. The peaks in the spectrum correspond to regions where TBP is present. Hence the coding regions are identified by putting a threshold to the spectrum or filtered energy sequence. The regions having energy above the threshold are considered as the protein-coding regions. Since the non-coding regions do not have a period-3 property, the energy in that region is low as demonstrated in Figure 3. It is interesting to note that the first coding region of 112 bp along positions 929-1039 has a weak TBP and the remaining four coding regions present high TBP. The spectral content method and anti-notch filter fail to detect properly that region, but the S-transform filtering approach catches up that region better than those two methods.

In order to have a comparison of the efficiency of these methods, the threshold percentiles from 1 to 99 are used on the measures of the individual methods for the identification of probable coding regions. Hence the statistical parameters, such as sensitivity, specificity and average accuracy, are calculated under the same conditions at different threshold values. The best result achieved in each method with the corresponding threshold value is listed in **Table 2**. The proposed method provides the best performance at a threshold of 85% with sensitivity 0.88, specificity 0.98 and average accuracy of 0.96. A comparative analysis of the average accuracy against the threshold values is shown in **Figure 4**. Furthermore, a comparison of the exon locations obtained from these three methods with those reported in NCBI database is listed in **Table 3**. It shows that the proposed method provides better discrimination between the exons and the introns compared to those offered by others. Again to assess the performance of the three methods, the receiver operating characteristic (ROC) curves are obtained. It is a representation of the prediction accuracy of separation of exons and introns in the gene. The ROC curve relates the true positive rate as a function of false positive rate for varying threshold values. The ROC curves for all the three methods are shown in **Figure 5**. The closer the ROC curve to a diagonal, the lesser effective the method at discrimi-

Table 2 Summary of the best performance (accuracy) of identification of coding regions in F56F11.4a using different methods

Method	Sensitivity	Specificity	Average accuracy	Threshold (%)
S-transform filter	0.88	0.98	0.96	85
DFT-based approach	0.82	0.86	0.85	81
Anti-notch filter	0.81	0.82	0.82	82

nating between exon and intron. The steeper the curve towards the vertical axis and then across, the better the method. A more precise way of evaluating the performance is to calculate the area under the ROC curve. The closer the area to 0.5, the poorer the method, and the closer to 1.0, the better the method. The area under the ROC curve for the S-transform filtering method is found to be 0.9288, and the same for the DFT and anti-notch filter methods are 0.8615 and 0.8369, respectively. Hence the proposed S-transform filtering method of exon prediction outperforms other methods as it offers the highest area under the curve.

We have also analyzed several DNA sequences from the benchmark dataset HMR195. The gene AF009614 has taken for demonstration and the power spectrum obtained from all the three methods is shown in **Figure 6**. The gene AF009614 has two exon regions at positions 1267-1639 and 3888-4513 in the sequence. From this figure, it is clearly elucidated that the proposed S-transform-based filtering method offers improved performance compared to its counterparts. We have also carried a classification experiment to compare the efficiency of the proposed method. From the HMR195 dataset, 50 sequences whose average exon length is greater than 200 bases are chosen for the experiment. Totally 222 coding sequences and 237 non-coding sequences are used in the study. The threshold percentile of 1-99 is used to discriminate the coding regions from the non-coding regions. Accordingly, the ROC curves by the three methods are shown in **Figure 7**. The areas under the ROC curve are also calculated. They are 0.8602, 0.8316 and 0.8094 for S-transform, DFT and anti-notch filter methods, respectively. Hence the S-transform based filtering method presents a better performance on the classification, thereby the superiority of the proposed method is assessed.

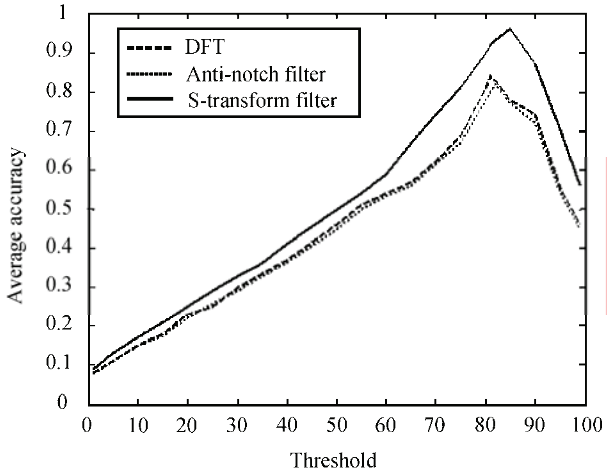

Figure 4 Average accuracy vs. threshold values of gene F56F11.4a.

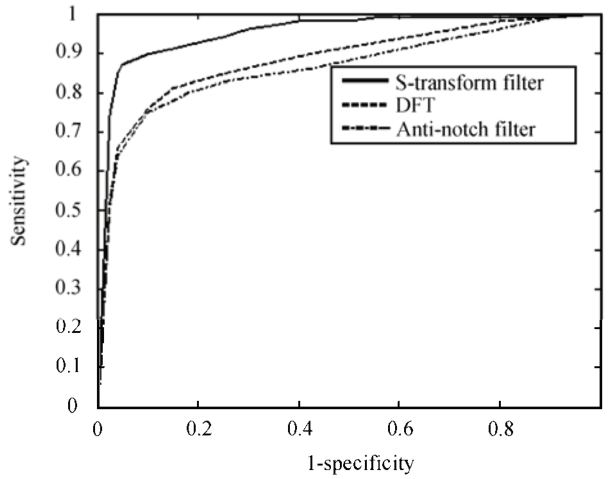

Figure 5 ROC curves obtained by DFT, anti-notch filter and S-transform filter from the gene F56F11.4a.

Table 3 Position comparison of the exons of F56F11.4a by different methods

Position in GenBank (NCBI)	DFT-based approach	Anti-notch filtering approach	S-transform filtering approach
929-1039 (110)	936-1169 (233)	942-1164 (222)	947-1037 (63)
2528-2857 (330)	2573-3005 (432)	2538-2956 (418)	2539-2908 (369)
4114-4377 (264)	4073-4432 (359)	4132-4462 (330)	4076-4409 (333)
5465-5644 (180)	5467-5658 (191)	5497-5672 (175)	5454-5644 (190)
7255-7605 (351)	7396-7806 (410)	7406-7728 (322)	7305-7597 (292)

Note: The length of the exons is shown in the braces.

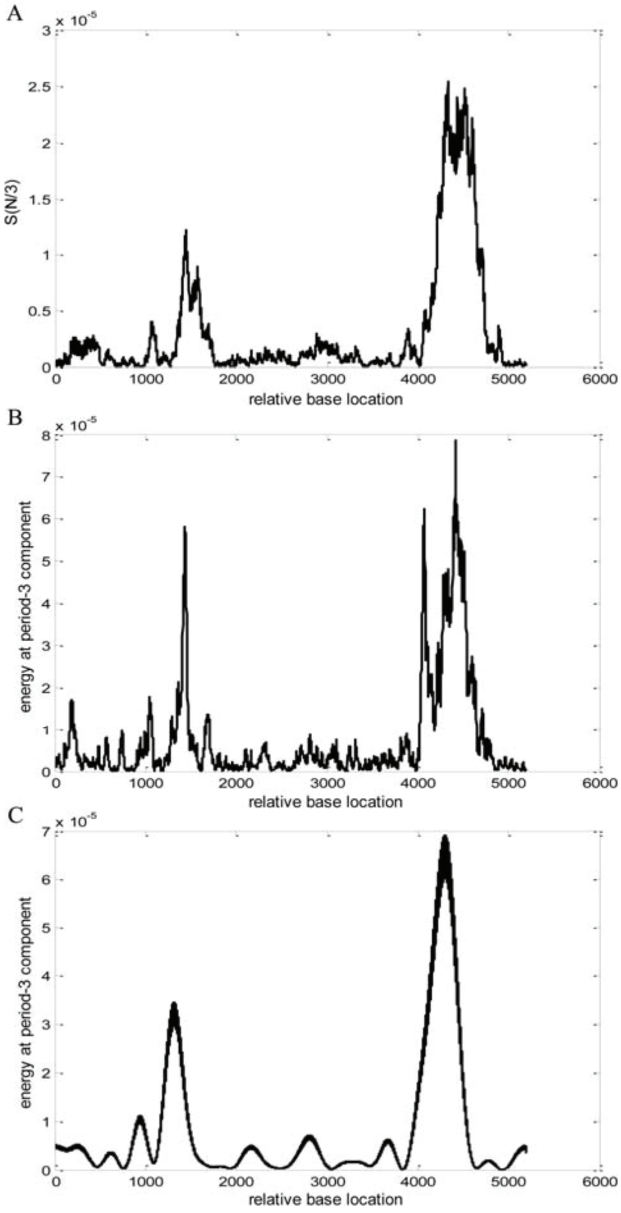

Figure 6 Comparison of the power spectra of the gene AF0099614 obtained by DFT (**A**), anti-notch filter (**B**) and S-transform filter (**C**).

Discussion

The existing exon identification methods employ a variety of biological information and coding techniques in association with many computational methods to predict the exon regions in DNA. Still the TBP pattern has been used as a basis to identify the coding regions. In this paper, we have introduced a new time-frequency filtering scheme based on the TBP for the identification of protein-coding regions. The S-transform method provides a pictorial view of the energy distribution of the frequencies with time, which helps in the analysis of the spectral varying signal. It gives a multi-resolution view of the signal so that distinct patches of periodic signal can be analyzed easily. It is a model-independent method that does not require any training sample to predict and is also independent of the window length constraint for proper computation of the spectra of coding regions. The multi-resolution analysis of the signal enables the proposed method to be effective for both small and large coding regions. Another aspect of the study is that the EIIP can be used as an efficient coding scheme for DNA sequence analysis. The proposed method is found to be robust against the background noise, which occurs due to long-range correlation of bases in the DNA sequence. Thus the coding regions can be better discriminated from the non-coding regions and thereby the accuracy of identification increases considerably. However, although the proposed method achieves better accuracy in the identification of the coding regions, it necessitates more computational effort. Another limitation of the S-transform method is that it provides low frequency resolution at higher frequencies and low time resolution at lower frequencies. This basically occurs due to the scaling nature of the Gaussian window during spectrum computation, which may affect the time-frequency filtering operation and also the accuracy. Hence improvement of the resolution of the spectrum can further improve the prediction accuracy.

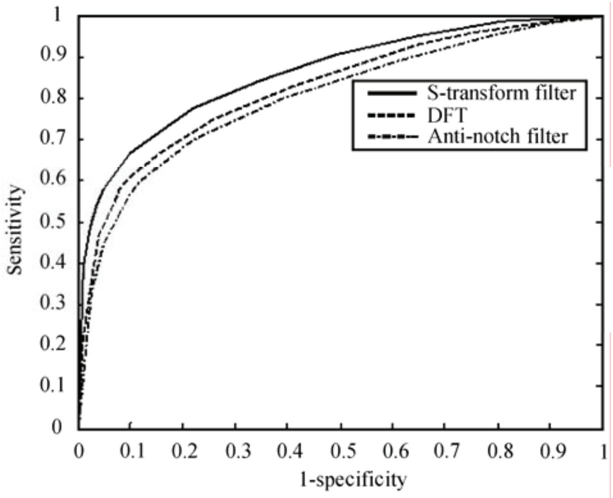

Figure 7 ROC curves obtained by three different methods (from 50 sequences of HRM195 dataset).

Conclusion

In this paper, an efficient time-frequency filtering approach is suggested for the identification of coding regions in the DNA sequence. The proposed method employs a multi-resolution approach to analyze both the small and large coding regions, and it does not depend on a prior window length as in case of Fourier methods. The performance of the proposed method is compared with the existing methods and the results show its superiority in identification of the coding regions. Thus it can be effectively used in DNA sequence analysis, such as promoter region identification and splice site detection.

Acknowledgements

The authors would like to thank Dr. Lalu Mansinha (University of Western Ontario, Canada) for his helpful discussions on S-transform technique.

Authors' contributions

SSS conceived and carried out the work, and drafted the manuscript. GP supervised the research, and participated in its design and coordination. Both authors have read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Fickett, J.W. and Tung, C.S. 1992. Assessment of protein coding measures. *Nucleic Acids Res.* 20: 6441-6450.

2 Fickett, J.W. 1996. The gene identification problem: an overview for developers. *Comput. Chem.* 20: 103-118.

3 Vaidyanathan, P.P. and Yoon, B.J. 2004. The role of signal-processing concepts in genomics and proteomics. *J. Franklin Inst.* 341: 111-135.

4 Tiwari, S., *et al.* 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.* 13: 263-270.

5 Tsonis, A.A., *et al.* 1991. Periodicity in DNA coding sequences: implications in gene evolution. *J. Theor. Biol.* 151: 323-331.

6 Gutierrez, G., *et al.* 1994. On the origin of the periodicity of three in protein coding DNA sequences. *J. Theor. Biol.* 167: 413-414.

7 Bernaola-Galvan, P., *et al.* 2000. Finding borders between coding and noncoding DNA regions by an entropic segmentation method. *Phy. Rev. Lett.* 85: 1342-1345.

8 Voss, R.F. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* 68: 3805-3808.

9 Chatzidimitriou-Dreismann, C.A. and Larhammar, D. 1993. Long-range correlations in DNA. *Nature* 361: 212-213.

10 Henderson, J., *et al.* 1997. Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.* 4: 127-141.

11 Ding, C.H. and Dubchak, I. 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17: 349-358.

12 Snyder, E.E. and Stormo, G.D. 1993. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural network. *Nucleic Acids Res.* 21: 607-613.

13 Eftestel, T., *et al.* 2006. Eukaryotic gene prediction by spectral analysis and pattern recognition techniques. In *Proceedings of the Seventh IEEE Nordic Signal Processing Symposium*, pp. 146-149. Reykjavik, Iceland.

14 Anastassiou, D. 2001. Genomic signal processing. *IEEE Sign. Proc. Mag.* 18: 8-20.

15 Fox, T.W. and Carreira, A. 2004. A digital signal processing method for gene prediction with improved noise suppression. *EURASIP J. Appl. Sign. Proc.* 2004: 108-114.

16 Datta, S. and Asif, A. 2005. A fast DFT based gene prediction algorithm for identification of protein coding regions. In *Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 653-656. York University, Toronto, Canada.

17 Chakravarthy, N., *et al.* 2004. Autoregressive modeling and feature analysis of DNA sequence. *EURASIP J. Appl. Sign. Proc.* 2004: 13-28.

18 Akhtar, M. 2005. Comparison of gene and exon prediction techniques for detection of short coding regions. *Int. J. Inf. Tech.* 11: 26-35.

19 Vaidyanathan, P.P. and Yoon, B.J. 2002. Digital filters for gene prediction applications. In *Proceedings of Asilomar Conference on Signals, Systems and Computers*, Vol. 1, pp. 306-310. Pacific Grove, USA.

20 Vaidyanathan, P.P. and Yoon, B.J. 2002. Gene and exon prediction using allpass-based filters. In *Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics*. Raleigh, USA.

21 Tuqan, J. and Rushdi, A. 2006. A DSP perspective to the period-3 detection problem. In *Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics*, pp. 53-54. University of California, Los Angeles, USA.

22 Rogic, S., *et al.* 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* 11: 817-832.

23 Silverman, B.D. and Linsker, R. 1986. A measure of DNA periodicity. *J. Theor. Biol.* 118: 295-300.

24 Zhang, R. and Zhang, C.T. 1994. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.* 11: 767-782.

25 Zhang C.T. and Wang J. 2000. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res.* 28: 2804-2814.

26 Nair, A.S. and Sreenadhan, S. 2006. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* 1: 197-202.

27 Rao, K.D. and Swamy, M.N.S. 2008. Analysis of genomics and proteomics using DSP techniques. *IEEE Trans. Circuits Syst.* 55: 370-378.

28 Cosic, I. 1994. Macromolecular bioactivity: is it resonant interaction between macromolecules?—Theory and applications. *IEEE Trans. Biomed. Eng.* 41: 1101-1114.

29 Sejdic, E., *et al.* 2009. Time-frequency feature representation using energy concentration: an overview of recent advances. *Digit. Signal Process.* 19: 153-183.

30 Qian, S. and Chen, D. 1996. Joint time-frequency analysis. *IEEE Signal Process Mag.* 16: 52-67.

31 Daubechies, I. 1990. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. Theory* 36: 961-1005.

32 Stockwell, R.G., *et al.* 1996. Localisation of the complex spectrum: the S transform. *IEEE Trans. Signal Process.* 44: 998-1001.

33 Rakovic, P., *et al.* 2006. Time-frequency signal processing approaches with applications to heart sound analysis. In *Proceedings of Computers in Cardiology*, pp. 197-200. Valencia, Spain.

34 Pinnegar, C.R. 2005. Time-frequency and time-time filtering with the S-transform and TT-transform. *Digit. Signal Process.* 15: 604-620.

35 Pinnegar, C.R. and Mansinha, L. 2003. The S-transform with windows of arbitrary and varying shape. *Geophysics* 68: 381-385.