PLOS ONE

RESEARCH ARTICLE

Exon prediction based on multiscale products of a genomic-inspired multiscale bilateral filtering

Xiaolei Zhang, Weijun Pan*

College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan, P.R. China

* panatc@sina.com

OPEN ACCESS

Citation: Zhang X, Pan W (2019) Exon prediction based on multiscale products of a genomic-inspired multiscale bilateral filtering. PLoS ONE 14 (3): e0205050. https://doi.org/10.1371/journal.pone.0205050

Editor: Vincenzo De Luca, University of Toronto, CANADA

Received: September 13, 2018

Accepted: March 5, 2019

Published: March 21, 2019

Data Availability Statement: The file HUMDZA2G (accession number D14034) of H. sapien is downloaded from the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). The data sets BG570 and HMR195 are available from the Institute of Microbial Technology Bioinformatics Center (http://www.imtech.res.in/raghava/genebench/datasets.html) for researchers who meet the criteria for access to confidential data. The sequences of the ENm001 and ENm004 datasets are available from NCBI GenBank according to the annotations of EGASP (http://genome.crg.es/datasets/egasp2005/).

Abstract

Multiscale signal processing techniques such as wavelet filtering have proved to be particularly successful in predicting exon sequences. Traditional wavelet predictor is domain filtering, and enforces exon features by weighting nucleotide values with coefficients. Such a measure performs linear filtering and is not suitable for preserving the short coding exons and the exon-intron boundaries. This paper describes a prediction framework that is capable of non-linearly processing DNA sequences while achieving high prediction rates. There are two key contributions. The first is the introduction of a genomic-inspired multiscale bilateral filtering (MSBF) which exploits both weighting coefficients in the spatial domain and nucleotide similarity in the range. Similarly to wavelet transform, the MSBF is also defined as a weighted sum of nucleotides. The difference is that the MSBF takes into account the variation of nucleotides at a specific codon position. The second contribution is the exploitation of inter-scale correlation in MSBF domain to find the inter-scale dependency on the differences between the exon signal and the background noise. This favourite property is used to sharp the important structures while weakening noise. Three benchmark data sets have been used in the evaluation of considered methods. By comparison with four existing techniques, the prediction results demonstrate that: the proposed method reveals at least improvement of 4.1%, 50.5%, 25.6%, 2.5%, 10.8%, 15.5%, 11.1%, 12.3%, 9.2% and 2.4% on the exons length of 1–24, 25–49, 50–74, 75–99, 100–124, 125–149, 150–174, 175–199, 200–299 and 300–300+, respectively. The MSBF of its nonlinear nature is good at energy compaction, which makes it capable of locating the sharp variations around short exons. The direct scale multiplication of coefficients at several adjacent scales obviously enhanced exon features while the noise contents were suppressed. We show that the non-linear nature and correlation-based property achieved in proposed predictor is greater than that for traditional filtering, which leads to better exon prediction performance. There are some possible applications of this predictor. Its good localization and protection of sharp variations will make the predictor be suitable to perform fault diagnosis of aero-engine.

Funding: This work was sponsored by the National Natural Science Foundation of China (Grant No. U1733203). This funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Recent advancement in high-throughput analysis, such as next-generation sequencing, has resulted in the development of computational techniques for the rapid prediction of exons in DNA sequences. Although great progress has been made in the development of exon prediction algorithms, the challenge of determining the lengths and locations of short exons urgently needs to be solved [1–3]. The main difficulty in predicting short exons is that the intrinsic properties, such as codon biases, are harder to determine [3,4]. To date, there is no consensus about the definition and classification of short exons. Saeys et al. thought that the exons with lengths of $<$200 base pair (bp) might be considered small [2]. Recently, two independent studies by Irimia et al. [5] in *Cell* and by Li et al. [6] in *Genome Research* defined one class of short exons called microexons and uncovered the features regulating the inclusion of these microexons. Irimia et al. reveal that the regulation of microexons (defined as exons with lengths of 3–15 bp) is highly dynamic during neuronal differentiation and the inclusion of these microexons can modulate the function of interaction domains of proteins involved in neurogenesis [5]. In another study, Li et al. demonstrate that microexons (defined as exons with lengths of $\leq$51 bp) exhibit a high level of sequence conservation and they may possess brain-specific functions [6]. Thus, knowledge pertaining to short exons in genomes is very important for understanding the functioning of proteins and the life processes. Therefore, the challenge of determining the lengths and locations of short exons urgently needs to be solved. In another work [7], we have briefly outlined the intrinsic advantages and limitations of the existing methods for predicting exons. In this paper, we focus on the development of a spectral analysis technique for finding exons in eukaryotic DNA sequences, as described below.

The discrete nature of DNA information has been driving a surging interest in the application of the principles of spectral analysis to develop efficient exon-prediction techniques. Spectral analysis techniques are attractive because they are easy to implement, entail reduced computational complexity, and mostly do not require any training of the genomic data [8–10]. In the spectral analysis of DNA sequences, the three-base periodicity (TBP) exhibited by exons is a good discriminator of coding potential. The determination of TBP due to codon usage bias is built upon the phenomenon that exon regions have a prominent power spectrum peak at frequency $f = 1/3$ [8,9]. Numerous advanced exon-finding algorithms have been developed by tracking the strength of TBP along a DNA sequence [3, 4, 7–21]. Such methodologies have a strong mathematical basis, including Fourier transform measures [8, 11–12], digital-filter-based methods [10,13], wavelet-based techniques [3,7,14–18] and other analysis tools [9]. Wavelets have proved highly successful in the manipulation and analysis of biomedical signals [22–28]. Among exon-finding methods, wavelet-based techniques are said to be distinctive. The examination of local variations in scale of the multiscale transform data of the sequence makes the wavelet predictor more powerful. Traditionally, the base idea of wavelet predictor, such as the modified Gabor-wavelet transform (MGWT) [14] and the wide-range wavelet window (WRWW) [18], is to computes a weighted sum of nucleotide values over a large neighbourhood at different scales. Although wavelet-based methods yield good predictions, they do not perform well in preserving the short exons and the exon-intron boundaries due to their linear nature. The multiscale bilateral filtering (MSBF) methods have been widely used in medical image processing field [29–31]. The nonlinear regularization of MSBF makes it an excellent solution for enhancing the high frequency structures and suppressing image noise. In this paper, we will follow the MSBF based strategy inspired from the one previously used in the analysis of image information to predict exons.

Our intuition is that nucleotides in the codon position $p(p = 1,2,3)$ are close to each other not only if they occupy nearby spatial locations but also if they have some similarity at the

reading frame p. For this purpose, we propose a genomic-inspired MSBF that can incorporate domain and similarity by means of multiplication. Like traditional wavelet predictor, a domain filtering named B-spline wavelet transform is designed to extract TBP by weighing nucleotide values with complex coefficients. Similarly, we define range filtering, which measures similarity by counting the sum of difference for variable sequence coverage. Another object of this paper is to investigate the inter-scale correlation (or multiscale products) information in MSBF domain and its application to exon prediction. We formulate the problem of investigating the correlated features in terms of the differences between exon and intron coefficients at two adjacent scales. We pursue this investigation which results from the HMR195 dataset by calculating the Jensen-Shannon divergence and the histogram distributions. Experimental results demonstrate that through MSBF and multiscale products, detection accuracy can be significantly improved with only a small loss in exon prediction. The proposed technique, termed multiscale products in MSBF domain (MP-MSBF), is more effective than locating exons directly from the linear filtering data, leading to superior exon prediction results.

2 Methodology

2.1 Numerical representation of a DNA sequence

The representation of DNA character strings into numerical sequences is the first step in DNA spectral analysis. In this paper, the paired-numerical representation [8] is introduced to map DNA characters (i.e., A, C, G, and T) into numeric values. A particular advantage of this representation is that it exploits the structural differences between exon and intron regions to facilitate the TBP extraction, in addition to reducing complexity. Eq (1) provides an example of this representation scheme for the short DNA fragment . . .*CTGCAGTGGT*. . .:

$$u = \{\ldots -1, 1, -1, -1, 1, -1, 1, -1, -1, 1 \ldots\}. \tag{1}$$

2.2 Genomic-inspired MSBF

To introduce our genomic-inspired MSBF, we first describe in Section A the domain filtering called B-spline wavelet transform. This wavelet function exhibits a higher degree of freedom for curve design, which can be adapted to analyse complex genome. In the next section, we first define a continuous representation of the average magnitude difference function (AMDF) inspired by Akhtar et al. [8], and a range filtering built with AMDF is designed to find certain information about nucleotide similarity in a specific codon position. Finally, the genomic-inspired MSBF is suggested for differentiating between intron noise and meaningful data.

A. Domain filtering. In this work, domain filtering given by B-spline windows are formulated. The B-spline window $\beta_m(t)$ of order m, which is time-limited in $[-T/2, T/2]$, is built as follows [32]:

$$\beta_m(t) = m^m \sum_{p=0}^{m} (-1)^p (t - (p - m/2)T/m)_+^{m-1} / p!(m-p)!, m = 1, 2, 3, \ldots, \tag{2}$$

where

$$(t - t_0) = \begin{cases} (t - t_0)^{m-1} & \text{if } t > t_0 \\ 0 & \text{if } t < t_0 \end{cases}. \tag{3}$$

Fig 1(A) plots $\beta_m(t)$ following Eqs (2) and (3).

To fully analyse the DNA sequences characterized by a specific periodicity, the task here is to extract the TBP at different scales while keeping the analysis frequency constant. From Eqs

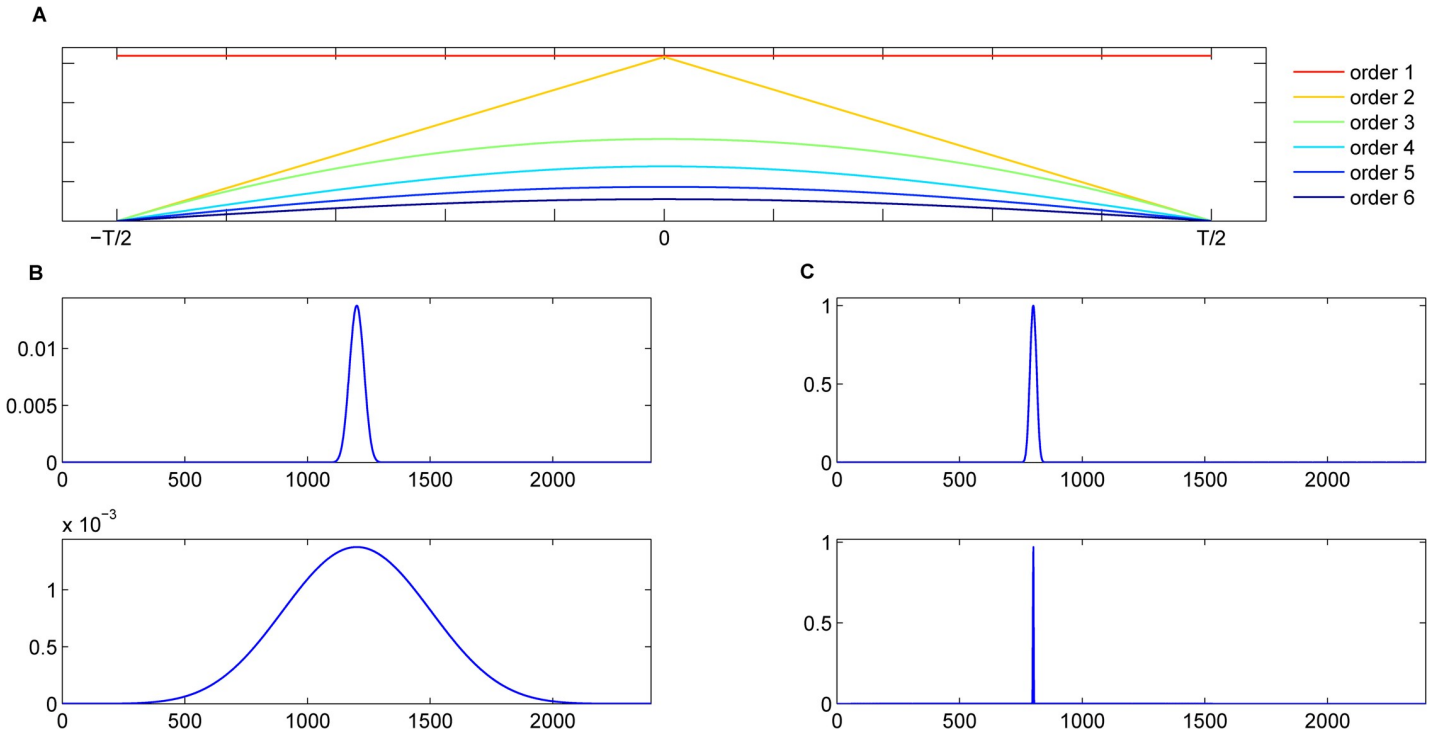

Fig 1. Examples of B-spline windows, and domain filter (order 6) with two different scales in the time and frequency domains. (A) Examples of B-spline windows; (B) Magnitude response of domain filter in the time domain; (C) Frequency response of domain filter.

https://doi.org/10.1371/journal.pone.0205050.g001

(2) and (3), our proposed domain filter of length L is defined as

$$\varphi_d(t,b,a) = \beta_m(t,b,a)e^{i\omega_0(t-b)}, \tag{4}$$

where $i^2 = -1$, $a>0$ is the scale (or dilation) parameter, b indicates the translation (or position) parameter, and $\omega_0 = L/3$ denotes the basic frequency. In Eq (4), the functions $\beta_m(t,b,a)$ are families generated from the base functions $\beta_m(t)$ by dilations and translations, i.e.,

$$\beta_m(t,b,a) = \frac{1}{\sqrt{a}}\beta_m\left(\frac{t-b}{a}\right). \tag{5}$$

Fig 1(B) and Fig 1(C) illustrate our domain filter with two different scales in the time and frequency domains. The proposed domain filtering of a signal u is given by

$$U_d(b,a) = \int u(t)\varphi_d(t,b,a)dt = \frac{1}{\sqrt{a}}\int u(t)\beta_m\left(\tfrac{t-b}{a}\right)e^{i\omega_0(t-b)}dt. \tag{6}$$

The domain filtering of Eq (6) measures the geometric distance between the center nucleotide and its neighbourhood.

In the case of domain filter, the length of the domain filter is 2400, and the scale parameter is set to 10 exponentially separated values between 1/60 and 1/6 for an input sequence. For practical purposes, the order of the B-spline function $\beta_m(t)$ is truncated to 6.

B. Range filtering. Before continuing to our genomic-inspired MSBF, we first use the AMDF to design a range filtering for measuring nucleotide similarity in a specific codon position. A continuous representation of AMDF for a signal u, as a function of the grid spacing τ_0,

is defined as

$$AMDF(t) = \frac{1}{L}\int_{t}^{t+L-1} |u(\tau) - u(\tau - \tau_0)|d\tau, \quad (7)$$

where L is equal to the window length of Eq (4), τ_0 is set to 3 for TBP. Before applying $AMDF$ to a DNA sequence, the authors in [8] suggest passing it first through a second-order resonant filter centered at frequency $2\pi/3$ [13].

For efficient implementation, a multiscale and sliding window will move along the filtered sequence to compute AMDF for the whole sequence. The complex envelope of $\varphi_d(t,b,a)$ given in Eq (4) is then used to calculate the window:

$$w(t, b, a) = |\varphi_d(t, b, a)|. \quad (8)$$

In other words, the window $w(t,b,a)$ is the magnitude response of $\varphi_d(t,b,a)$ in time domain. From Eqs (7) and (8), the proposed range filtering for a signal u can be formulated as follows:

$$U_r(b, a) = \frac{1}{L}\int[\int_{t}^{t+L-1} |(u(\tau) - u(\tau - \tau_0))w(\tau, b, a)|d\tau]dt. \quad (9)$$

The range filtering of Eq (9) measures the radiometric distance between the center nucleotide and its neighbourhood.

Finally, the expressions given in Eqs (6) and (9) are used to design our genomic-inspired MSBF of a signal u, having the non-linear property:

$$U(b, a) = U_d(b, a) \cdot U_r(b, a). \quad (10)$$

Given a DNA sequence of length N, the projection of the MSBF coefficients onto the position axis is defined as a function of $b(b = 0,1,\ldots,N-1)$.

2.3 Multiscale products

Several exon-finding techniques take advantage of traditional wavelet transform to filter short exons with small scales and long exons with large scales. This approach implies that they do not exploit the dependencies between adjacent scales. To explore the MSBF inter-scale correlations we multiply the adjacent MSBF sub-bands to distinguish intron noise from meaningful data while preserving the sharp variations of short exons. The core idea behind the multiscale products method is based on our research (see **Section 3.4**): namely, for DNA sequences represented by MSBF, the multiscale transform coefficients related to intron noise are less correlated across scales than the coefficients associated with exon signals.

Let $U(b,a_j)$ be the MSBF of a signal u at the scale $a_j(j = 1,2,\ldots,J)$ and the position b. The multiscale products (or inter-scale correlation) $MP_j(b)$ of the MSBF contents at two adjacent scales is defined as

$$MP_j(b) = |U(b, a_j)| \cdot |U(b, a_{j+1})|, \; j = 1, 2, \ldots, J - 1. \quad (11)$$

With the observation of experimental results, we can imagine that multiplying the MSBF at adjacent scales would amplify exon structures and dilute noise (see **Section 3**).

2.4 Multiscale products of multiscale bilateral filtering

Our multiscale products of multiscale bilateral filtering (MP-MSBF) for exon prediction is described briefly in Table 1. The input DNA sequence of length N is referred to as u.

Table 1. Exon predictor algorithm using the MP-MSBF technique.

1. Convert an input DNA sequence into the numerical sequence u using the paired-numerical representation.
2. Apply the MSBF to the whole sequence. The transform of the numerical sequence is given by

$$U(b, a_j) = U_d(b, a_j) \cdot U_r(b, a_j),$$

where $a_j (j = 1,2,\ldots,J)$ is the scale parameter and b denotes the nucleotide position along the DNA sequence.

3. Take $U(a_j,b)$ as an input and perform the multiscale products to obtain the filtered sequence $MP_j(b) (j = 1,2,\ldots,J-1)$.
4. Compute the spectrum of the numerical DNA sequence: $S(b) = \Sigma_j |MP_j(b)|^2$.
5. Project the obtained spectrum onto the position axis, which is defined as a function of b:

$$S_p(b) = S(b), b = 0, 1, \ldots, N-1.$$

https://doi.org/10.1371/journal.pone.0205050.t001

3 Results and discussion

3.1 Data resources

To evaluate and compare the performance of the proposed MP-MSBF with that of other methods, the two benchmark data sets BG570 [33] and HMR195 [34] have been considered. Furthermore, we conduct an additional classification experiment using 29 genes of the ENm001-004 data set (part of EGASP) [35] (see **S1 File** for detailed information on these sequences). Table 2 summarizes the features of the considered data sets.

3.2 General setting

In this section, we first conduct an experiment to establish a comprehensive analysis of the inter-scale correlation of the differences between exon and intron coefficients. Next, we present experiments in exon prediction using the proposed method. For comparison, MP-MSBF presents comparable performance to that of four popular existing methods: the paired and weighted spectral rotation measure (PWSR) [8], the MGWT [14], the fast Fourier transform plus empirical mode decomposition (FFTEMD) [11] and the WRWW [18]. To evaluate the general performances of these measures, the TBP data for each DNA sequence considered have been normalized with values between 0 and 1.

3.3 Evaluation metrics

To investigate the inter-scale correlation of the differences between exon and intron sequences, the distance criterion of Jensen-Shannon (JS) divergence [36] is adopted. In probability theory and statistics, the JS divergence is a method of measuring the similarity between two probability distributions. The JS divergence is a convenient divergence measure for our purpose because it is symmetric and bounded between 0 and 1. The distance between two probability

Table 2. Statistics of the test data sets.

Dataset	Species	Genes	Length	Exons	Average length of exons (bp)	Proportion of exons/introns
BG570	Vertebrate	570	2,892,149	2,649	168	15.37% / 84.63%
HMR195	Mammalian	195	1,383,720	948	208	14% / 86%
EGASP	Human	29	2,425,886	323	167	2.22% / 97.78%

https://doi.org/10.1371/journal.pone.0205050.t002

vectors $\boldsymbol{P}$ and $\boldsymbol{Q}$ in terms of the JS divergence is defined as

$$JS(\boldsymbol{P},\boldsymbol{Q}) = \frac{1}{2}KL(\boldsymbol{P},\boldsymbol{M}) + \frac{1}{2}KL(\boldsymbol{Q},\boldsymbol{M}), \tag{12}$$

where $\boldsymbol{M} = (\boldsymbol{P}+\boldsymbol{Q})/2$ and KL is the Kullback-Leibler divergence,

$$KL(\boldsymbol{P},\boldsymbol{M}) = \sum_{l} p(l)\log_2\left(\frac{p(l)}{m(l)}\right). \tag{13}$$

With a set of results obtained by running a predictor on a test data set, the true positive (TP), true negative (TN), false negative (FN) and false positive (FP) counts can be determined. Using these counts, the performances of various methods in handling exons of different lengths are measured in terms of the approximate correlation (AC) [33]

$$AC = \left(\frac{1}{4}\left(\frac{TP}{TP+FN}+\frac{TP}{TP+FP}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right)-0.5\right)\times 2. \tag{14}$$

To evaluate the general performance of the method under consideration, the receiver operating characteristic (ROC) curve [37] is used to explore the effects on *sensitivity* and *specificity*. The sensitivity and specificity are given by

$$Sensitivity = \frac{TP}{TP+FN}, \tag{15}$$

$$Specificity = \frac{TN}{TN+FP}. \tag{16}$$

The area under an ROC curve (AUC) can be used as an indicator of prediction performance.

3.4 Inter-scale correlation analysis

The coefficients of the input DNA sequences obtained from the multiresolution decomposition include exon-structure information together with intron noise. The general purpose of inter-scale correlation analysis is to investigate the dependency information on the differences between exon and intron coefficients. We apply the schemes proposed in this paper to analyse the correlation for a large number of exon and intron regions.

Fig 2 shows the prediction plots of the sequence HUMDZA2G locus (*AZGP1* gene) of *Homo sapiens* (GenBank accession number D14034) using MSBF and its inter-scale correlation (or multiscale products) at different scales. The sequence HUMDZA2G locus (*AZGP1* gene) contains four exons at positions 5322–5388, 9329–9589, 12907–13182 and 14052–14335. The peaks corresponding to the exon regions of the original data appear much stronger in Fig 2(B) than those in Fig 2(A). The results demonstrate that inter-scale correlation can suppress intron noise while retaining more exon details. Fig 3 compares the prediction results of the sequence HUMDZA2G locus (*AZGP1* gene) using the tested methods. Our MP-MSBF algorithm identified the localized peaks better and located the short coding sequence (exon 1) more accurately.

Herein, the JS divergences are employed to investigate whether the coefficients related to introns are less correlated across scales than the coefficients associated with exons. This distance criterion has been applied in genome comparison [38], bioinformatics [39] and protein surface comparison [40]. Table 3 summarizes the JS divergences of the MSBF coefficients between two adjacent scales, $a_{j,j+1}$ (j = 1,2,. . .,9), for the exon and intron nucleotides of the HMR195 data set. The results of Table 3 reveal that the JS divergences of exons are smaller

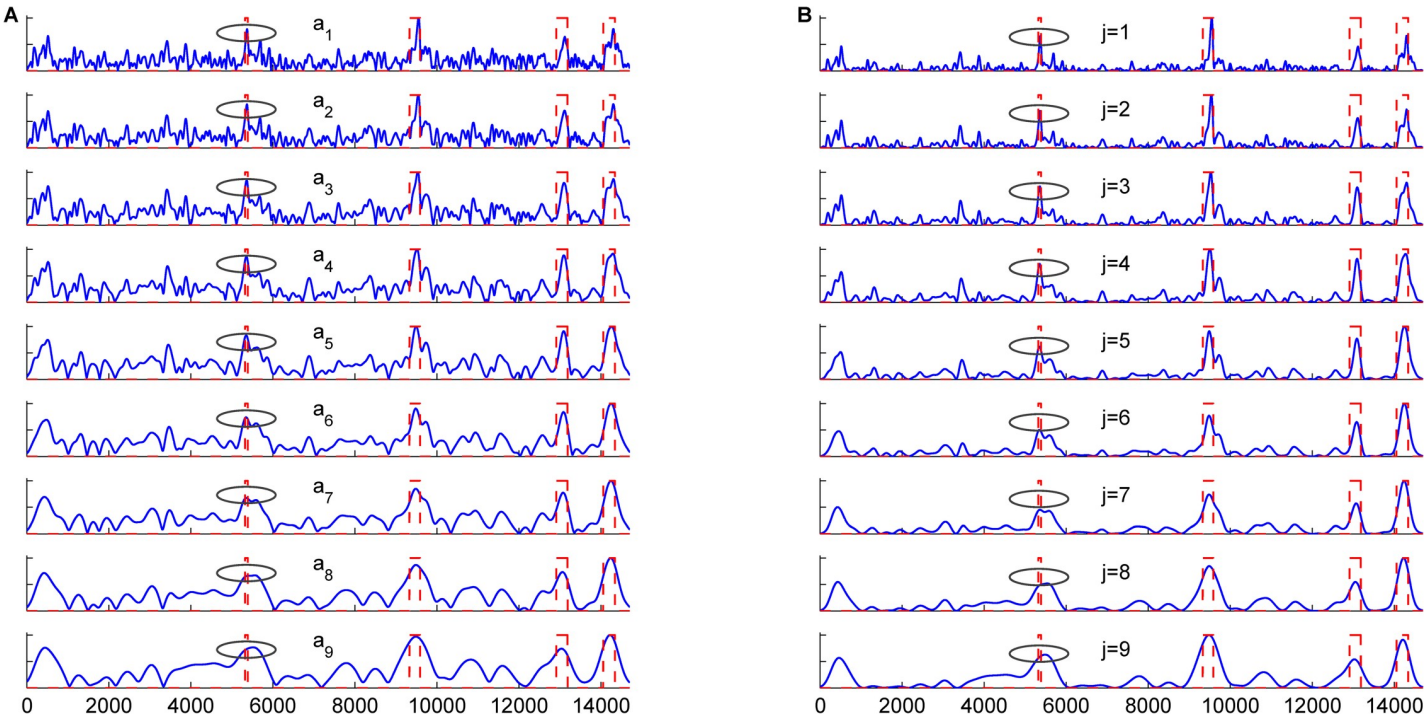

Fig 2. Prediction plots for sequence HUMDZA2G locus (*AZGP1* gene) at different scales. The abscissa axes of all the plots represent the relative base positions, the actual locations of the exons are marked with rectangles in red dashed lines. Part (A) shows the MSBF result; and (B) shows the result of inter-scale correlation.

https://doi.org/10.1371/journal.pone.0205050.g002

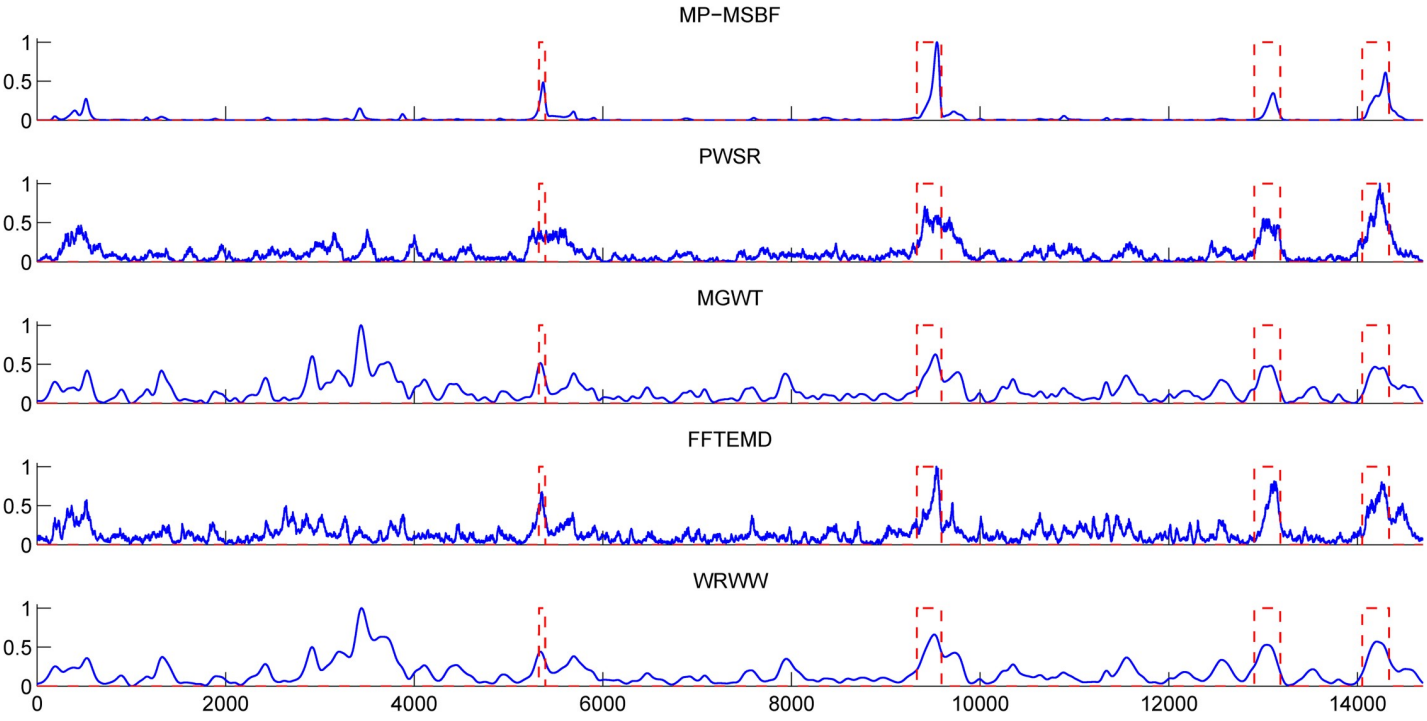

Fig 3. Prediction results for the sequence HUMDZA2G locus (*AZGP1* gene) using the considered methods. The abscissa axes of all the plots represent the relative base positions, and the actual locations of exons are marked with rectangles in red dashed lines.

https://doi.org/10.1371/journal.pone.0205050.g003

Table 3. JS divergence of MSBF coefficients between two adjacent scales.

Regions	JS divergence of MSBF coefficients between consecutive scales								
	$a_{1,2}$	$a_{2,3}$	$a_{3,4}$	$a_{4,5}$	$a_{5,6}$	$a_{6,7}$	$a_{7,8}$	$a_{8,9}$	$a_{9,10}$
Exon	0.0030	0.0028	0.0032	0.0057	0.0083	0.0089	0.0041	0.0035	0.0033
Intron	0.0085	0.0130	0.0189	0.0220	0.0206	0.0138	0.0102	0.0133	0.0178

https://doi.org/10.1371/journal.pone.0205050.t003

than those of introns at consecutive scales, while there is a difference of one order of magnitude between the exon and intron regions at the last eight consecutive scales. This property can assist in discriminating exon features from introns in the multiscale transform domain.

To further justify our assumption, histograms with fitted distributions are calculated for the exon and intron nucleotides of HMR195 at different scales. Fig 4(A) and Fig 4(B) give the distributions of exon and intron nucleotides using the MSBF and inter-scale correlation, respectively. This result indicates that the most relevant exon information represented by the correlation at each scale is captured by large-valued coefficients, whereas the intron information is captured by a large number of small-valued coefficients. Fig 5 clearly illustrates that the distance between the exon and intron curves obtained from inter-scale correlation is greater than that obtained from MSBF. In other words, the MSBF coefficients of the exon sequences have a strong correlation on various decomposition scales, whereas the MSBF coefficients of noise are weakly correlated. These plots justify our assumption.

3.5 Performance evaluation on benchmark data sets

Exons have significant functional constraints, and their length plays an important role in splice site selection. Rogic's evaluation work [34] stated, "*These constraints have shaped the exon length distribution quite differently from geometric distribution. The length distribution depends on the exon type*." In our analysis, we grouped exons into ten ranges of exon lengths, namely, (0,25), [25,50), [50,75), [75,100), [100,125), [125,150), [150,175), [175,200), [200,300) and [300,300+). We thought the exons of these ranges are relatively short and long in length. The best accuracies achieved by the tested methods are calculated in terms of the AC values for each group of exons. Fig 6(A) and Fig 6(B) depict the experimental results obtained from various methods using the HMR195 and BG570 data sets, respectively. The MP-MSBF exhibits good accuracies in these ten ranges. In Fig 6(A), MP-MSBF presents results close to those of MGWT in the ranges (0,25) and WRWW in the ranges [300,300+), while it exceeds the performance of other methods in the other ranges. The results of Fig 6(B) show that the performance of the MP-MSBF method is close to those of FFTEMD at the range (0,25) and MGWT at the range [75,100), while it outperforms the performance of the other methods at the other ranges of exon lengths. Similar results obtained with the sequences in the ENm00-004 data set are shown in Fig 6(C); however, no exons of length <25 occur in these sequences. The MP-MSBF exhibits good accuracies in these nine ranges and presents results close to those of FFTEMD at the ranges [25,75) and WRWW in the ranges [100,125), [200,300) and [300,300+), while it slightly exceeds the performance of other methods in the other ranges.

Table 4 summaries the performances of various methods for exons using the BG570, HMR195 and ENm001-004 data sets. By comparison with the PWSR, MGWT, FFTEMD and WRWW methods, the prediction results show that: our MP-MSBF exhibits at least improvement of 4.1%, 50.5%, 25.6%, 2.5%, 10.8%, 15.5%, 11.1%, 12.3%, 9.2% and 2.4% on the exons of the ranges (0,25), [25,50), [50,75), [75,100), [100,125), [125,150), [150,175), [175,200), [200,300) and [300,300+), respectively.

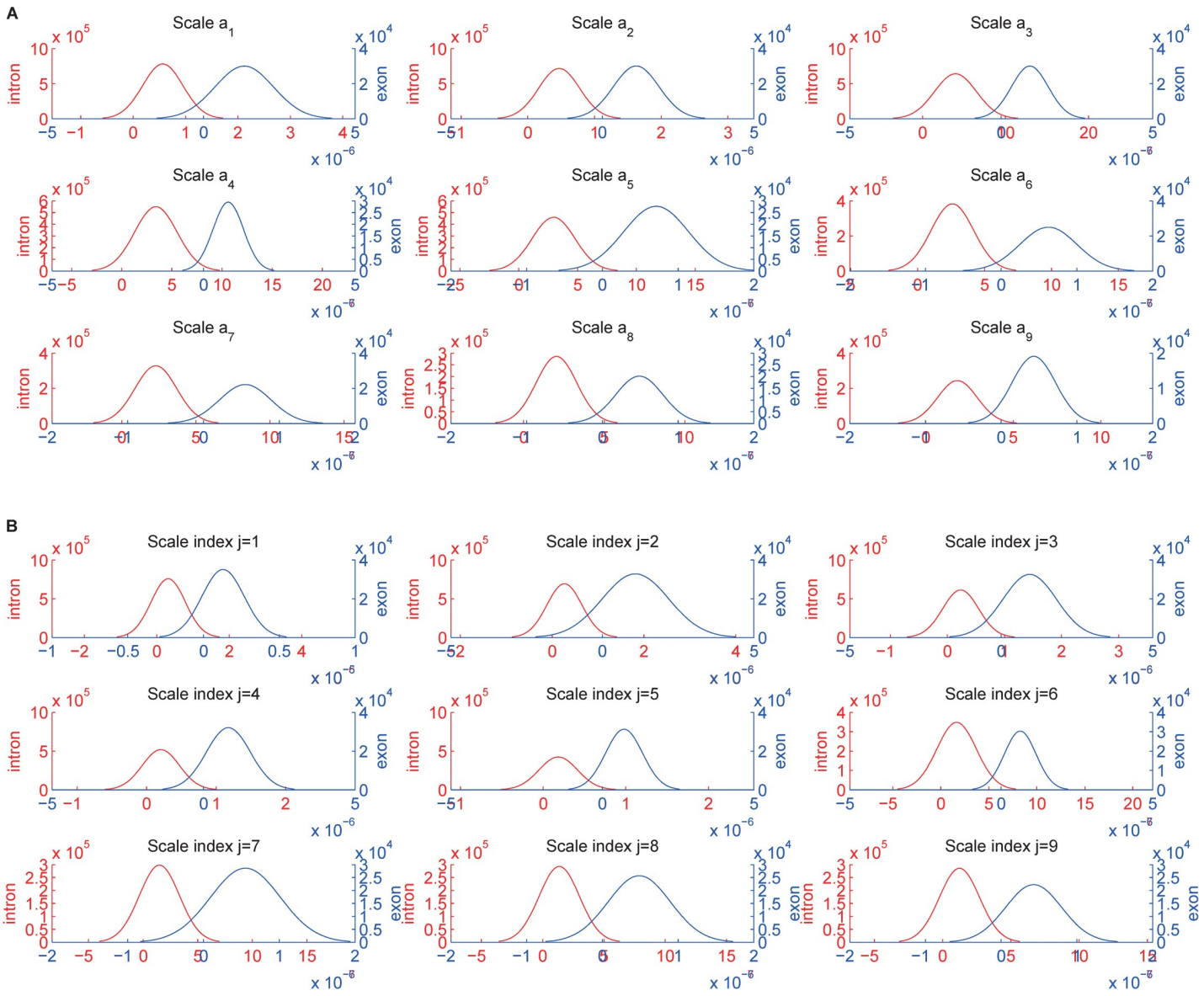

Fig 4. Histogram distributions at different scales for MSBF and inter-scale correlation applied to HMR195. For all the plots, blue lines represent exons, red lines indicate introns, the abscissa axes represent the magnitude values, and the ordinate axes represent the number of coefficients. Part (A) shows the MSBF result; and (B) shows the result of inter-scale correlation.

https://doi.org/10.1371/journal.pone.0205050.g004

An additional classification experiment on all sequences of considered data sets is designed to assess the general performance of our proposed technique and other methods. Fig 7 presents the ROC curves obtained from the different methods tested in this experiment. The MP-MSBF method has higher prediction accuracy than its counterparts. Our MP-MSBF method consistently exhibits higher prediction accuracy than its counterparts for exons that are either relatively short or long in length.

3.6 Summary

In this work, we have introduced a new, robust and efficient method to predict exons in eukaryotes. Unlike some prediction techniques that detect exons directly by linear filtering,

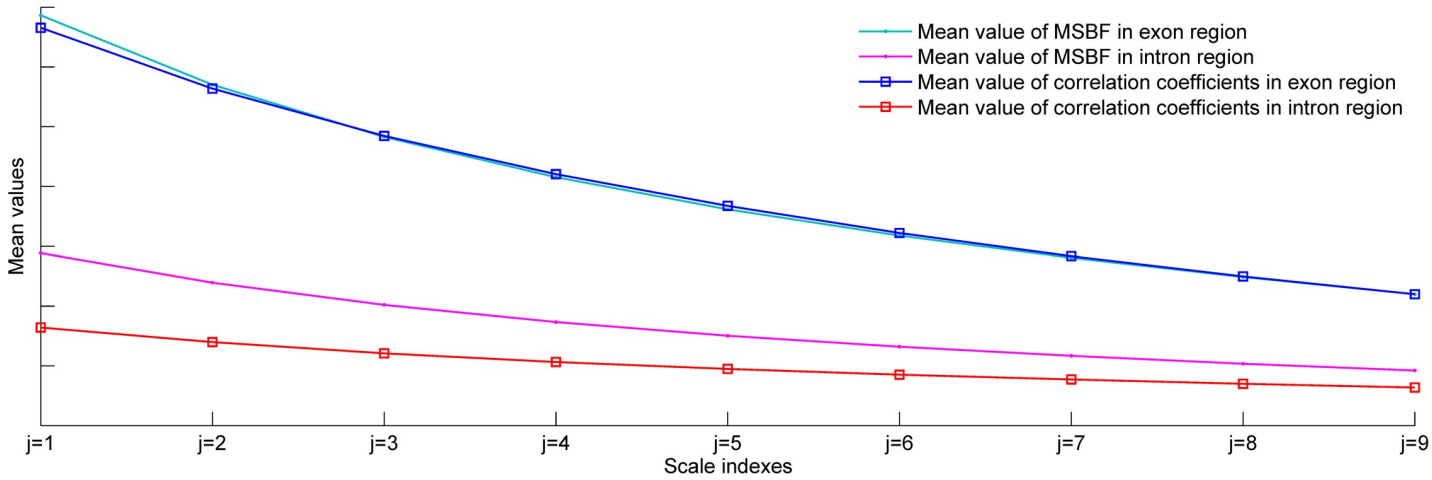

Fig 5. Mean values of histogram distributions at different scales for MSBF and inter-scale correlation applied to HMR195.

https://doi.org/10.1371/journal.pone.0205050.g005

the proposed scheme incorporates a genomic-inspired multiscale bilateral filtering and its inter-scale dependencies and then applies these features to better differentiate exon structures from background noise. The first key concept of our method is its nonlinear nature, which exploits geometric distance in the spatial domain and nucleotide similarity in the range. The second is that this technique compacts the energy of the exons into coefficients with large amplitudes and spreads the energy of the introns over a large number of coefficients with small amplitudes. This phenomenon has led to improved results with respect to exon preservation and noise suppression. The proposed MP-MSBF method requires neither prior

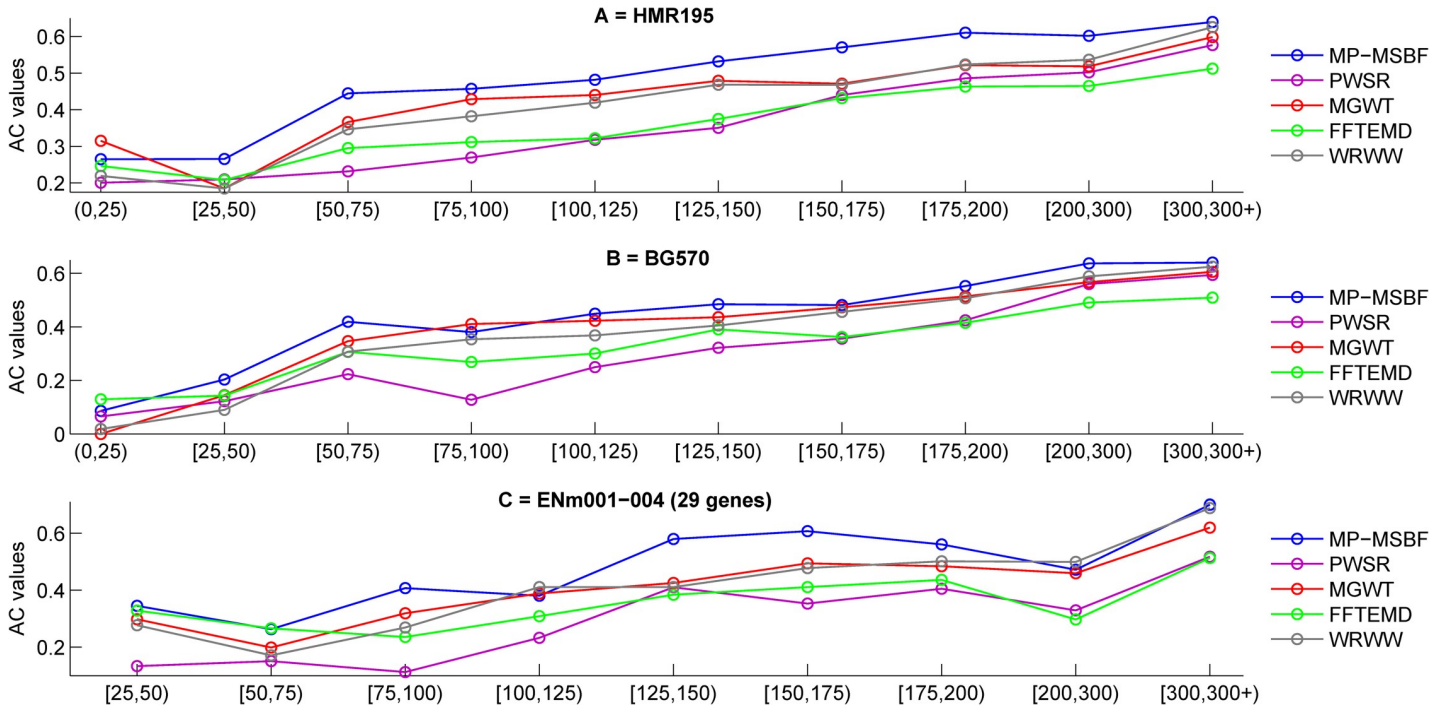

Fig 6. Plots of approximate correlation (AC) for considered data sets with various methods applied to exons in length ranges. For all the plots, the ordinate axes denote the ranges of exon lengths.

https://doi.org/10.1371/journal.pone.0205050.g006

Table 4. Best performances obtained from considered methods for exons using the BG570, HMR195 and ENm001-004 data sets.

Methods	Approximate coefficient (AC)									
	(0,25)	[25,50)	[50,75)	[75,100)	[100,125)	[125,150)	[150,175)	[175,200)	[200,300)	[300,300+)
MP-MSBF	0.205	0.292	0.466	0.459	0.505	0.566	0.563	0.623	0.665	0.682
PWSR	0.117	0.161	0.239	0.197	0.299	0.378	0.414	0.477	0.559	0.612
MGWT	0.117	0.181	0.371	0.448	0.456	0.490	0.507	0.549	0.585	0.635
FFTEMD	0.197	0.194	0.333	0.333	0.352	0.439	0.424	0.475	0.510	0.547
WRWW	0.103	0.172	0.341	0.407	0.429	0.481	0.505	0.555	0.609	0.666

https://doi.org/10.1371/journal.pone.0205050.t004

information nor training models for exon prediction, and so it can be applied to analyse unknown and novel genomes. It should be noted that all five methods considered here tend to have low accuracy in predicting microexons shorter than 25 bp (there were only 121) and microexons with lengths of 25–49 bp (there were only 227) as shown in Table 4. A possible explanation for this phenomenon is that these microexons are too short to be efficiently spliced in vivo without special splicing activation sequences [41]. In other words, the length of these microexons is too short to be clearly distinguished from surrounding noncoding regions. For almost all the methods, the accuracies slowly rise with the length of annotated exons between 75 and 300+ nucleotides. Although not good for exons shorter than 50 bp, the results obtained from our method are acceptable. Our MP-BSBF should encourage further development of existing methods in prediction of microexons.

4 Conclusion

Exons encode the biochemical processes and information involved in the pathway from DNA to proteins. In genomic sequence analysis, exon prediction based on the annotated sequences

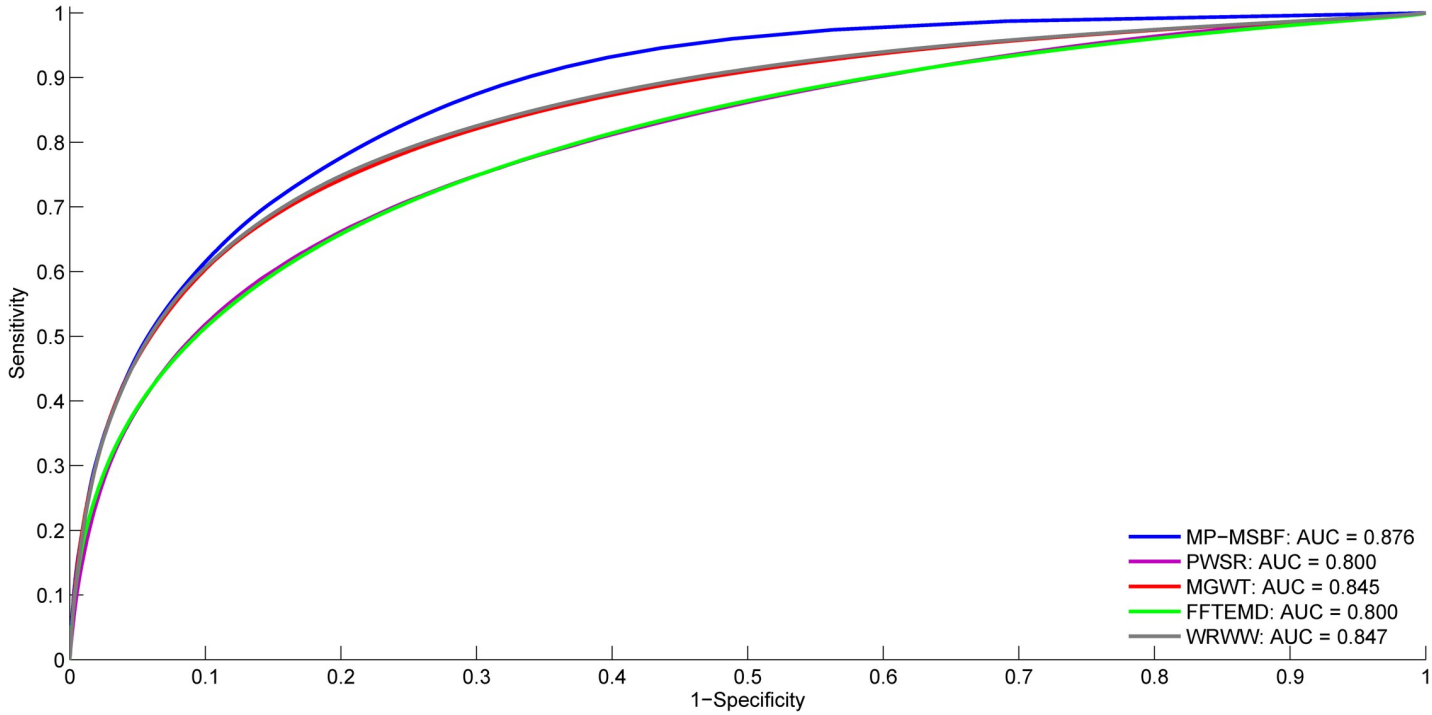

Fig 7. ROC plots of tested methods using the BG570, HMR195 and ENm001-004 data sets.

https://doi.org/10.1371/journal.pone.0205050.g007

in the online databases is an important problem. For exon prediction, extracting the relevant features of short coding sequences is a major task because the subtle features of short exons are obscured by the strong presence of background noise. In practice, spectral analysis is an important tool for the discovery of interesting patterns and structures in exon data. In this paper, we present a new exon-finding spectral analysis method that overcomes some of the shortcomings of current predicting techniques. The MP-MSBF predictor takes advantage of the nonlinear filtering and the dependency information between scales, which makes it capable of short exon prediction. We see some possible applications of this predictor. The correlation-based property and nonlinear nature of this technique allow the selection of a characteristic frequency from surrounding noise and thereby makes it possible to offer good localization and protection of sharp variations for locating hot spots in proteins and performing fault diagnosis of aero-engine.

Supporting information

S1 File. Sequences of the ENm001-004 data set used for the analyses presented in this paper. Detailed information on these sequences. (ZIP)
(ZIP)

Author Contributions

Conceptualization: Xiaolei Zhang.

Data curation: Xiaolei Zhang.

Formal analysis: Xiaolei Zhang.

Funding acquisition: Weijun Pan.

Investigation: Xiaolei Zhang, Weijun Pan.

Methodology: Xiaolei Zhang.

Project administration: Xiaolei Zhang, Weijun Pan.

Resources: Xiaolei Zhang, Weijun Pan.

Software: Xiaolei Zhang.

Supervision: Xiaolei Zhang, Weijun Pan.

Validation: Xiaolei Zhang, Weijun Pan.

Visualization: Xiaolei Zhang.

Writing – original draft: Xiaolei Zhang.

Writing – review & editing: Weijun Pan.

References

1. Wu Y, Liew AW-C, Yan H, Yang M. Classification of short human exons and introns based on statistical features. Phys Rev E. 2003; 67(6):061916.
2. Saeys Y, Rouzé P, Van de Peer Y. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. Bioinformatics. 2007; 23(4):414–20. https://doi.org/10.1093/bioinformatics/btl639 PMID: 17204465
3. Jiang R, Yan H. Studies of spectral properties of short genes using the wavelet subspace Hilbert–Huang transform (WSHHT). Physica A: Statistical Mechanics and its Applications. 2008; 387 (16):4223–47.

4. Jiang R, Yan H. Segmentation of short human exons based on spectral features of double curves. Int J Data Min Bioinform. 2008; 2(1):15–35. PMID: 18399326
5. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014; 159(7):1511–23. https://doi.org/10.1016/j.cell.2014.11.035 PMID: 25525873
6. Li YI, Sanchez-Pulido L, Haerty W, Ponting CP. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. 2015; 25(1):1–13. https://doi.org/10.1101/gr.181990.114 PMID: 25524026
7. Zhang X, Shen Z, Zhang G, Shen Y, Chen M, Zhao J, et al. Short Exon Detection via Wavelet Transform Modulus Maxima. PloS One. 2016; 11(9):e0163088. https://doi.org/10.1371/journal.pone.0163088 PMID: 27635656
8. Akhtar M, Epps J, Ambikairajah E. Signal processing in sequence analysis: advances in eukaryotic gene prediction. IEEE J Sel Top Signal Process. 2008; 2(3):310–21.
9. Marhon SA, Kremer SC. Gene prediction based on DNA spectral analysis: a literature review. J Comput Biol. 2011; 18(4):639–76. https://doi.org/10.1089/cmb.2010.0184 PMID: 21381961
10. Ramachandran P, Lu WS, Antoniou A. Filter-based methodology for the location of hot spots in proteins and exons in DNA. IEEE Trans Biomed Eng. 2012; 59(6):1598–609. https://doi.org/10.1109/TBME.2012.2190512 PMID: 22410955
11. Zhang W-F, Yan H. Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences. Pattern Recogn. 2012; 45(3):947–55.
12. Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A fast algorithm for exonic regions prediction in DNA sequences. J Med Signals Sens. 2013; 3(3):139–49. PMID: 24672762
13. Vaidyanathan P, Yoon B-J. The role of signal-processing concepts in genomics and proteomics. J Franklin Inst. 2004; 341(1):111–35.
14. Mena-Chalco JP, Carrer H, Zana Y, Cesar RM Jr. Identification of protein coding regions using the modified Gabor-wavelet transform. IEEE/ACM Trans Comput Biol Bioinform. 2008; 5(2):198–207. https://doi.org/10.1109/TCBB.2007.70259 PMID: 18451429
15. Shakya DK, Saxena R, Sharma SN. An adaptive window length strategy for eukaryotic CDS prediction. IEEE/ACM Trans Comput Biol Bioinform. 2013; 10(5):1241–52. https://doi.org/10.1109/TCBB.2013.76 PMID: 24384711
16. Zhang X, Zhao J, Xu W, editors. Identification of eukaryotic exons using empirical mode decomposition and modified Gabor-wavelet transform. Proceedings of the 33rd Chinese Control Conference; 2014 July 28–30; Nanjing, China. IEEE Institute of Electrical and Electronics Engineers Inc; 2014.
17. Zhao J, Zhang X, Xu W. Prediction of Eukaryotic Exons via the Singularity Detection Algorithm. Curr Bioinform. 2014; 9(4):389–401.
18. Marhon S, Kremer S. Prediction of protein coding regions using a wide-range wavelet window method. IEEE/ACM Trans Comput Biol Bioinform. 2016; 13(4):742–53. https://doi.org/10.1109/TCBB.2015.2476789 PMID: 26415183
19. Ahmad M, Jung LT, Bhuiyan AA. A biological inspired fuzzy adaptive window median filter (FAWMF) for enhancing DNA signal processing. Comput Meth Prog Bio. 2017; 149:11–7.
20. Zhang G, Zhang X, Pan G, Yu Y, Chen Y. Improved prediction of short exons via multiscale products. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2017 Oct 14–16; Shanghai, China. IEEE Institute of Electrical and Electronics Engineers Inc; 2017.
21. Zhang X, Zhang G, Yu Y, Pan G, Deng H, Shi X, Jiao Y, Wu R, Chen Y. Multiscale Products in B-spline Wavelet Domain: A New Method for Short Exon Detection. Curr Bioinform. 2018; 13(5): 553–563.
22. Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafa Y, Thermes C. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Phys Rep. 2011; 498(2):45–188.
23. Audit B, Baker A, Chen CL, Rappailles A, Guilbaud G, Julienne H, Goldar A, d'Aubenton-Carafa Y, Hyrien O, Thermes C, Arneodo A. Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. Nat Protoc. 2013; 8(1):98–110. https://doi.org/10.1038/nprot.2012.145 PMID: 23237832
24. Butler W E. Wavelet brain angiography suggests arteriovenous pulse wave phase locking. PloS One. 2017; 12(11):e0187014. https://doi.org/10.1371/journal.pone.0187014 PMID: 29140981
25. Chen D, Wan S, Xiang J, et al. A high-performance seizure detection algorithm based on Discrete Wavelet Transform (DWT) and EEG. PloS One. 2017; 12(3):e0173138. https://doi.org/10.1371/journal.pone.0173138 PMID: 28278203

26. Rajagopal S., Tharcismariapushpam I., Improved Algorithm for the Location of CPG Islands in Genomic Sequences Using Discrete Wavelet Transforms, Curr Bioinform. 12 (9) (2017) 57–65.
27. Toplak T, Palmieri B, Juanes-García A, et al. Wavelet Imaging on Multiple Scales (WIMS) reveals focal adhesion distributions, dynamics and coupling between actomyosin bundle stability. PloS One. 2017; 12(10):e0186058. https://doi.org/10.1371/journal.pone.0186058 PMID: 29049414
28. Wang XH, Jiao Y, Li L. Mapping individual voxel-wise morphological connectivity using wavelet transform of voxel-based morphology. PloS One. 2018; 13(7):e0201243. https://doi.org/10.1371/journal.pone.0201243 PMID: 30040855
29. Hu J, Li S. The multiscale directional bilateral filter and its application to multisensor image fusion. Inform Fusion. 2012; 13(3):196–206.
30. Samala RK, Chan HP, Lu Y, Hadjiiski L, Wei J, Sahiner B, et al. Computer-aided detection of clustered microcalcifications in multiscale bilateral filtering regularized reconstructed digital breast tomosynthesis volume. Med Phys. 2014; 41(2):021901. https://doi.org/10.1118/1.4860955 PMID: 24506622
31. Lu Y, Chan HP, Wei J, Hadjiiski LM, Samala RK. Multiscale bilateral filtering for improving image quality in digital breast tomosynthesis. Med Phys. 2015; 42(1):182–95. https://doi.org/10.1118/1.4903283 PMID: 25563259
32. Toraichi K, Kamada M, Itahashi S, Mori R. Window functions represented by B-spline functions. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1989; 37(1):145–7.
33. Burset M, Guigó R. Evaluation of gene structure prediction programs. Genomics. 1996; 34(3):353–67. https://doi.org/10.1006/geno.1996.0298 PMID: 8786136
34. Rogic S, Mackworth AK, Ouellette FB. Evaluation of gene-finding programs on mammalian sequences. Genome Res. 2001; 11(5):817–32. https://doi.org/10.1101/gr.147901 PMID: 11337477
35. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, et al. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 2006; 7 Suppl 1:S2.1–31.
36. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory. 1991; 37 (1):145–51.
37. Pencina M, D'Agostino R Sr, D'Agostino R Jr, Vasan R. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med. 2008; 27(2):157–72. https://doi.org/10.1002/sim.2929 PMID: 17569110
38. Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci U S A. 2009; 106(8):2677–82. https://doi.org/10.1073/pnas.0813249106 PMID: 19188606
39. Itzkovitz S, Hodis E, Segal E. Overlapping codes within protein-coding sequences. Genome Res. 2010; 20(11):1582–9. https://doi.org/10.1101/gr.105072.110 PMID: 20841429
40. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol. 2003; 325(2):377–87. PMID: 12488102
41. Dominski Z, Kole R. Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol. 1991; 11(12):6075–83. PMID: 1944277