Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 166–173

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Original Research

A Modified Statistically Optimal Null Filter Method for Recognizing Protein-coding Regions

Lei Zhang [1,*], Fengchun Tian [1], Shiyuan Wang [2]

[1] *College of Communication Engineering, Chongqing University, Chongqing 400044, China*
[2] *School of Electronic and Information Engineering, Southwest University, Chongqing 400715, China*

Received 15 September 2011; revised 4 February 2012; accepted 21 February 2012
Available online 19 June 2012

Abstract

Computer-aided protein-coding gene prediction in uncharacterized genomic DNA sequences is one of the most important issues of biological signal processing. A modified filter method based on a statistically optimal null filter (SONF) theory is proposed for recognizing protein-coding regions. The square deviation gain (SDG) between the input and output of the model is used to identify the coding regions. The effective SDG amplification model with Class I and Class II enhancement is designed to suppress the non-coding regions. Also, an evaluation algorithm has been used to compare the modified model with most gene prediction methods currently available in terms of sensitivity, specificity and precision. The performance for identification of protein-coding regions has been evaluated at the nucleotide level using benchmark datasets and 91.4%, 96%, 93.7% were obtained for sensitivity, specificity and precision, respectively. These results suggest that the proposed model is potentially useful in gene finding field, which can help recognize protein-coding regions with higher precision and speed than present algorithms.

Keywords: Gene prediction; Biological signal processing; Protein-coding region; Square deviation gain

Introduction

Recognition of protein-coding regions has attracted much attention in recent years. Currently, different kinds of methods for locating protein-coding regions have been proposed. Conventional techniques for recognizing exons of DNA sequences include intelligent methods based on neural networks [1], hidden Markov models (HMMs) based on statistical theory [2–4], and correlation function methods [5]. Markov chain based models perform well in gene-findings for genome sequence analysis [6] and a better Markov model, which relies on a number of training gene datasets for accurate model parameters such as the first, second and fifth-order Markov models, has been well developed in comparison with other algorithms using Z-curve [7]. However, the computational speed of such models is cost-ineffective in the training process. Besides, the prior information is overconsidered in modeling. Thus, further development of more convenient and simple algorithms with acceptable accuracy is beneficial to genome sequence studies especially in the investigation of eukaryote genomes.

In recent years, signal processing approaches have attracted significant attention in research of genomic sequences and genome structures, which may identify hidden periodicity and features that cannot be revealed easily by conventional statistical methods. In DNA sequences, protein-coding regions typically show a periodic character of three bases, which cannot be found in intergenic regions and introns in eukaryotes. Previous digital signal processing methods were based on the property of period-3 and include the discrete Fourier transform (DFT) [8–11], short-time discrete Fourier transform (STDFT) with a sliding window [12], lengthen-shuffle DFT based on the format of the Z-curve [13] and EPND method with DNA walk

[*] Corresponding author.
E-mail: leizhang@cqu.edu.cn (Zhang L).

http://dx.doi.org/10.1016/j.gpb.2012.02.001

sequences [14]. In addition, the band-pass digital filters, which are centered at $2\pi/3$, have also been proposed to predict protein-coding regions. These include single infinite impulse response (IIR) anti-notch filter using lattice structure [15] and multistage filters to suppress the background 1/f noise [16]. Guigo divides gene prediction methods into model-dependent and model-independent methods [17]. Model-dependent methods depend on a priori known genomic information of organisms, while model-independent methods do not. Model-independent methods, such as modified Garbor transform (MGT) [18], DSP methods [19,20], extended Kalman filters based on symbolic dynamics [21], and a time-frequency filtering technique based on S-transform [22], can be used to identify unknown protein-coding regions of DNA sequences. However, most existing algorithms may be useful in the recognition of DNA sequences which are longer but the accuracy of recognition may be affected for shorter sequences. Recently, an exon detection algorithm using statistically optimal null filters (SONFs) has been proposed in comparison with the DFT algorithm for shorter sequences [23] and showed its feasibility in gene prediction for shorter DNA sequences. SONF, which is closely related to the Kalman filter, reduced the modeling complexity without requiring the solution of nonlinear equations of the Ricatti type which is essential in computing the gain of the Kalman filter [24]. SONF has been widely used in the seizure detection field [25]. Effectiveness and lower complexity of computing with SONF inspire us to explore SONF in-depth for detecting more favorable characteristics of genomes.

For clarity, the outline of this paper has been shown as follows. First, we apply Z-curve representation to map DNA sequences into digital sequences. Second, we illustrate the basic principle of the improved model and the square deviation (SD) gain (SDG) amplification method was used to suppress the non-coding signal which is viewed as $1/f$ noise in this paper. The complete recursive iteration algorithm is also described, and the global procedure frame is given. Then, we describe the benchmark gene datasets, and present the prediction of coding regions with Class I and Class II amplification using the proposed algorithm. Furthermore, an evaluation measurement is performed for comparison with other gene prediction methods using the F56F11.4 sequence, HMR195 datasets and human β-globin gene, respectively. Finally, a conclusion of this paper is presented. This paper addresses the challenges in the locations of longer DNA sequences and shorter DNA sequences, respectively, using SONF without any training datasets, which is different from the Markov chain based models that require DNA sequence length and their a-priori biological information.

Methods

The improved model and structure are used to detect protein-coding regions. Also, a SDG amplification method is applied to suppress the non-coding regions. The recursive algorithm is illustrated and the global program for implementation is presented in detail. A block diagram of gene prediction model is shown in **Figure 1**.

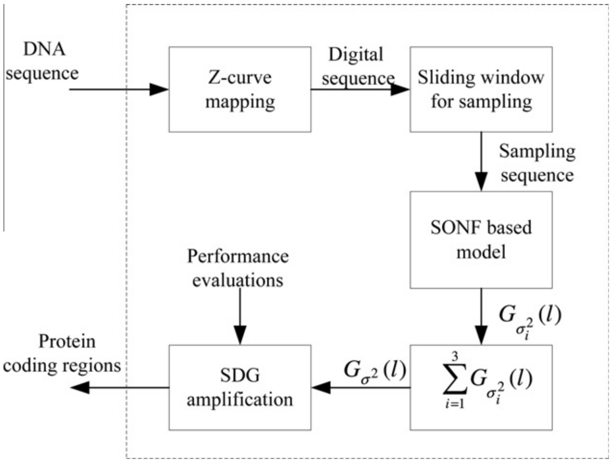

Figure 1 The schematic data flow block diagram of the proposed gene prediction model structure

Digital mapping of DNA sequence

DNA sequence digitalization is the first stage in genome analysis. We describe a three-dimensional curve representation called the Z-curve to reconstruct each base [26].

Considering a DNA sequence with N bases, we calculate the cumulative numbers of the bases A, C, G and T, respectively beginning from the 1st base to the n-th base. We then obtain four positive integers A_n, C_n, G_n and T_n. The Z-curve is constructed by a group of nodes P_n ($n = 1, 2, \ldots, N$), whose coordinates are illustrated by the following x_n, y_n, and z_n [26]:

$$\begin{cases} x_n = 2(A_n + G_n) - n, \\ y_n = 2(A_n + C_n) - n, \quad n = 0, 1, \ldots, N \\ z_n = 2(A_n + T_n) - n, \end{cases} \tag{1}$$

where the initial values $A_0 = C_0 = G_0 = T_0 = 0$ and $x_0 = y_0 = z_0 = 0$. Also, we define that

$$\begin{cases} \Delta x_n = x_n - x_{n-1}, \\ \Delta y_n = y_n - y_{n-1}, \quad n = 1, 2, \ldots, N \\ \Delta z_n = z_n - z_{n-1}, \end{cases} \tag{2}$$

Thus, we know that a DNA sequence can be decomposed into three digital sequences (consisting of 1 or -1), which represents the distribution of purine/pyrimidine type, amino/keto type and strong/weak hydrogen bonds type along the DNA sequences, respectively [13].

To detect protein-coding region, a sliding window with a width of M samples is applied in our model, where M should be determined by the maximum exon length in protein-coding regions. To obtain a new digital DNA sequence, the window is then moved by one base overlapping for every sample interval until all the bases are embedded in the window.

Modified filtering

The basic theory of instantaneous matched filter (IMF) has been presented previously [27] and a general description of the SONF, which combines the maximum output signal-to-noise ratio (SNR) with the minimum mean square error (MMSE) criteria, is also given [24]. The modified filter method designed based on IMF and SONF is shown as follows.

Consider a DNA sequence with a length of N shown as follows:

$$x(k) = d(k) + n(k) \tag{3}$$

where $n(k)$ is zero-mean white Gaussian noise, and the desired signal $d(k)$ is represented as:

$$d(k) = V^T\phi(k) \tag{4}$$

where V is the random variable, $\phi(k)$ is the known basis function, and $v(k)$ is the output of IMF which can be expressed as:

$$v(k) = \sum_{k=0}^{N} x(k)\phi(k) \tag{5}$$

For the least optimization of IMF output, we scale $v(k)$ by $\phi(k)/c(k)$, then the desired signal estimate can be shown by:

$$\hat{d}(k) = [Vc(k) + n_0'(k)]\phi(k)/c(k) = V\phi(k) + n_0(k) \tag{6}$$

where $n_0(k)$ is white noise, and the ultimate form of $d(k)$ is illustrated as:

$$\hat{d}(k) = d(k) + n_0(k) \tag{7}$$

To determine the optimal filter, we scale the output $v(k)$ of IMF by an unknown function $\lambda(k)$, the final output $y(k)$ of IMF is presented as:

$$y(k) = v(k)\lambda(k) \tag{8}$$

where $y(k)$ is also called the estimate of $d(k)$, and the output error $\varepsilon(k)$ of the filter is illustrated as:

$$\varepsilon(k) = x(k) - y(k) = x(k) - v(k)\lambda(k) \tag{9}$$

By using (3) the output error $\varepsilon(k)$ can also been represented by:

$$\varepsilon(k) = d(k) + n(k) - v(k)\lambda(k) \tag{10}$$

For an ideal null filter, $\varepsilon_{ideal}(k) = n(k)$, and the error function of the filter becomes:

$$\begin{aligned}\varepsilon_\lambda(k) &= \varepsilon_{ideal}(k) - \varepsilon(k)\\ &= n(k) - [d(k) + n(k)] + v(k)\lambda(k)\end{aligned} \tag{11}$$

Consider the MSE criteria, with respect to the input SNR, the optimal post-IMF scaling function $\lambda_{opt}(k)$ can be written as:

$$\lambda_{opt}(k) = \phi(k)/[q(k) + 1/SNR] \tag{12}$$

where $q(k)$ is shown as follows:

$$q(k) = \sum_{i=0}^{k} \phi(k)^2 \tag{13}$$

Thus, the power of the input noise should be small enough (*i.e.* $SNR \to \infty$), then the scaling function is rewritten as:

$$\lambda'_{opt}(k) = \phi(k)/q(k) \tag{14}$$

The detailed structure of the model is illustrated in **Figure 2**.

SDG amplification model

Non-coding regions may obscure the coding regions in prediction such that, the border information of coding regions cannot be identified accurately [28]. To suppress the non-coding regions ($1/f$ noise), an SDG amplification method is proposed to enhance the SDG of coding regions which is recognized as our feature object of the coding regions. A related suppressing method has been introduced in which a quadratic window operation is performed [28]. The window can effectively suppress the non-coding regions while preserving the coding regions so that the coding regions can be easily recognized. However, a different window length is needed for different DNA sequences. To calculate the SDG of DNA sequence segment, we first design a SDG function as the weight scales of original output which is similar to the signal boosting method [29]. We define the SDG of coding regions as follows

$$R_r = R_{r-1} + \eta(G_r - R_{r-1}) \tag{15}$$

where η is a smaller positive value (*i.e.*, $\eta = 0.2 \pm 0.05$) which is equal to smooth coefficient to control the sensitivity of algorithm. The SDG of non-coding regions is illustrated as follows:

$$Q_r = \begin{cases} \mu \cdot Q_{r-1}, & if\ Q_{r-1} \leqslant R_r \\ R_r, & if\ Q_{r-1} > R_r \end{cases} \tag{16}$$

where μ should be slightly greater than 1 to control the attenuation velocity of noise level. Therefore, the final object value after Class I amplification is described as:

$$\bar{G}_r = \Psi_r^2 G_r \tag{17}$$

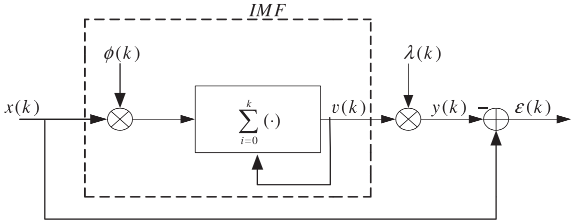

Figure 2 The structure of the improved filter model
The part with dashed line denotes the block of instantaneous matched filter (IMF).

where the gain function $\Psi_r = R_r/Q_r$, G_r is the original SDG of coding regions. The operation on the original filter results denotes the Class I amplification, while the Class II amplification denotes the same operation on the Class I amplification results so that more signal of non-coding region will be suppressed.

Implementation of recursive algorithm

Before the implementation, we first define SD of signal X as:

$$\sigma^2 = E\{[X - E(X)]^2\} \tag{18}$$

where $E(X)$ denotes the expected value of signal X.

The SDG between the input and output of iteration algorithm, G_{σ^2}, is recognized as the resultant object of the detection of coding regions which is defined as:

$$G_{\sigma^2} = \sigma_o^2/\sigma_i^2 \tag{19}$$

Considering the period-3 property of protein-coding regions, we determine the dimensions of model as follows:

$$\begin{aligned} \lambda(k) &= [\lambda_1(k)\ \lambda_2(k)\ \lambda_3(k)]^T \\ \phi(k) &= [\phi_1(k)\ \phi_2(k)\ \phi_3(k)]^T \\ v &= [v_1\ v_2\ v_3]^T \end{aligned} \tag{20}$$

The basis functions with a desired period-3 property primarily perform a good forecast property [23]. In this paper, we have considered several different selections for the parameter c of $\phi_i(k)$ (c is an uncertain constant, $i = 1, 2, 3$) for generality. The three orthogonal basis functions have been illustrated as follows:

$$\begin{aligned} \phi_1(k) &= (c\,0\,0\,c\,0\,0\,c\,0\,0\ldots) \\ \phi_2(k) &= (0\,c\,0\,0\,c\,0\,0\,c\,0\ldots) \\ \phi_3(k) &= (0\,0\,c\,0\,0\,c\,0\,0\,c\ldots) \end{aligned} \tag{21}$$

Combining the introduced theory [24], the complete recursive algorithm has been presented as follows:

$$v(k) = v(k-1) + x(k)\phi(k) \tag{22a}$$

$$P(k) = P(k-1) - P(k-1)\phi(k)\phi^T(k)P(k-1)/[1 + \phi^T(k)P(k-1)\phi(k)] \tag{22b}$$

$$\lambda(k) = P(k)\phi(k) \tag{22c}$$

$$y(k) = v^T(k)\lambda(k) \tag{22d}$$

$$\varepsilon(k) = x(k) - v^T(k)\lambda(k) \tag{22e}$$

The recursive update formula for the gain matrix $P(k)_{3\times 3}$ in this paper originated from the matrix lemma shown below:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1} \tag{23}$$

where $A = P(k-1)^{-1},\ B = \phi(k),\ C = I,\ D = \phi(k)^T$.

The gain matrix $P(k)$ is initialized as matrix I and $v(k)$ is initialized as $v(0) = (0,0,0)^T$ for iterations; where I is the identity matrix.

```
1. Map DNA sequence into three digital sequences $x_1(n)$, $x_2(n)$ and $x_3(n)$ using Z-curve
2. for $l$=1, do
3.     $l$++
4.     get sub-sequences $x_{1,l}(k)$, $x_{2,l}(k)$ and $x_{3,l}(k)$ by sliding the window by one base
5.     Initialization of filter
6.     for $k$=1 to $M$, do
7.         apply recursive filtering procedures into $x_{1,l}(k)$, $x_{2,l}(k)$ and $x_{3,l}(k)$, respectively
8.     end for
9.   Apply the SDG algorithm illustrated in subsection SDG amplification model
10.    if $l$+351 is smaller than $N$
11.    Return to 2 for next iteration
12.    else iteration terminated
13.    end if
14. end for
15. Apply the SDG amplification algorithm to suppress the non-coding regions
16. end
```

Figure 3 Implementation diagram of the modified algorithm

Combining our DNA representation of Z-curve and the SDG amplification operation with the recursive algorithm, we summarize the brief global iteration implementation program for locating protein coding regions as shown in Figure 3.

Results and discussion

Gene datasets

To evaluate the performance of the improved filter model on the detection of protein-coding regions, we apply the iteration implementation program to the gene F56F11.4 on the *Caenorhabditis elegans* chromosome III which contains five known coding exons in positions 928–1039, 2528–2857, 4114–4377, 5465–5644, and 7255–7605 (GenBank accession number AF099922 [21].

In this work, one benchmark dataset from the mammalian organism HMR195 dataset has also been considered. HMR195 is a dataset of 195 sequences with exactly one complete either single-exon or multi-exon gene. HMR195 has the following characteristics: (1) the ratio of human:mouse:rat sequences is 103:82:10, (2) the mean length of the sequences in the set is 7096 bp, (3) the number of single-exon genes is 43, and the number of multi-exon genes is 152, (4) the average number of exons per gene is 4.86, (5) the mean exon length is 208 bp, the mean intron length is 678 bp and the mean coding length of a gene is 1015 bp (~330 amino acids), and (6) the proportion of coding sequence in this dataset is 14%, of the intronic sequence 46% and of the intergenic DNA 40%.

Application of the modified model to gene datasets

The proposed algorithm is applied to three gene sequences with known fragments of exons, respectively. First, for the long sequence F56F11.4, the sliding window width is set as

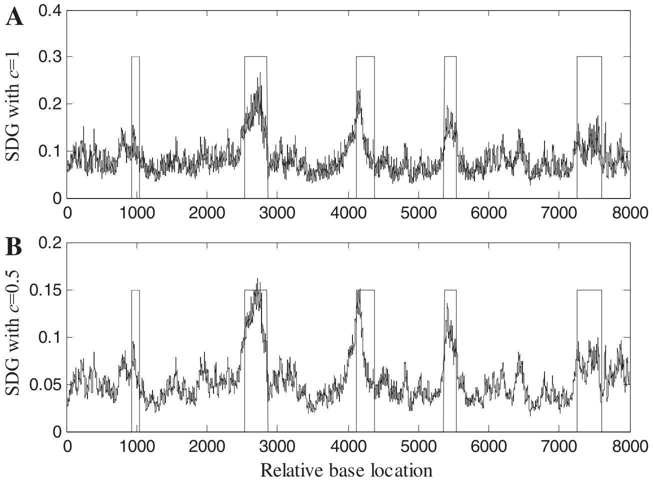

Figure 4 Identification of coding regions on F56F11.4
The output SDG of model for $c = 1$ and $c = 0.5$ was shown in A and B, respectively. The binary dot lines illustrate the true coding exons regions for visualization. The vertical axis shows the SDG, and the horizontal axis shows the relative base location.

$M = 351$ according to the maximum exon length and the sliding window step is 1 bp. **Figure 4** illustrates the prediction performance of the model combined with SDG algorithm under the condition that parameter c equals to 1 and 0.5, respectively. The locations of peak values are recognized as the predicted exon areas using our method. The peak values show that the SDG is larger in coding regions. It is consistent with the theory that the SNR between the coding regions (signal) and the non-coding regions (noise) is large [28]. When comparing Figure 4A with B, we observe that the plots become smoother with the decreasing of parameter c which is very similar to the smooth filter.

Figure 5 illustrate the SDG after Class I and Class II amplification, respectively. We can see that the non-coding regions are suppressed effectively, and the accurate border locations of coding regions are visible. From **Figure 6**, we can see that the non-coding regions almost tend to zero after Class II amplification. Therefore, we can say that the SDG amplification method is effective in recognition of coding regions.

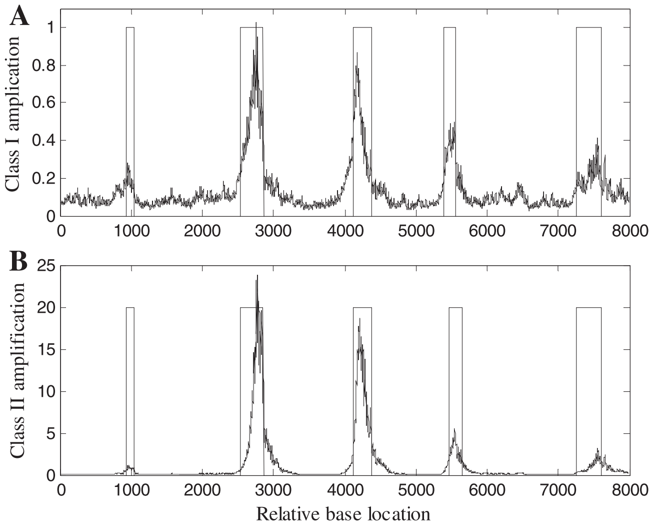

Figure 5 Identification of coding regions on F56F11.4 with Class I (A) and Class II (B) amplification

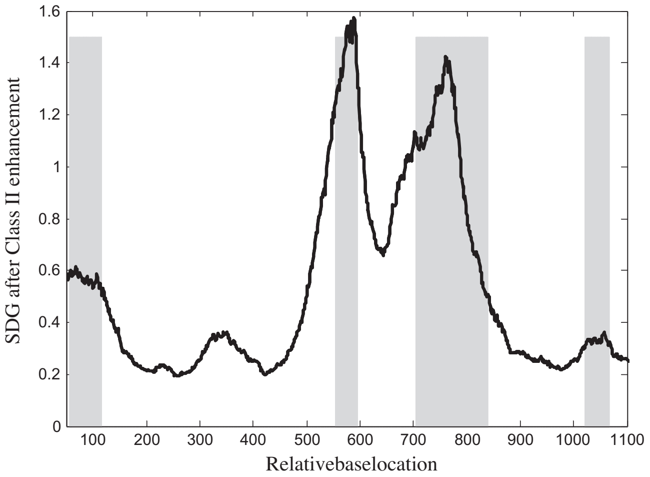

Figure 6 Recognition of the No. 5 mammalian sequence in HMR195 dataset after Class II enhancement using the improved model
The gray regions denote the relative physical positions of CDS features.

Figure 6 illustrates the performance of the extracted 1100 bps from the No. 5 mammalian sequence in datasets HMR195. The true coding regions (57-117, 554-595, 706-839 and 1022-1067) have been shown intuitively with CDS feature in GenBank.

Model evaluation

To evaluate the validity of the proposed recognition model in *C. elegans*, a modified evaluation scheme is performed on the basis of the previous publications [18,30]. A threshold *th* percent smaller than the SDG is viewed as the non-coding regions, and set to zero similarly as described previously [18]. The threshold value *th* is in the range between 1 and 99 to predict the borders of coding regions for calculating sensitivity (Sn), specificity (Sp) and the precision (P). In this paper, the best threshold *th* is 83. **Figure 7** illustrates the nucleotide-level measures of prediction borders. The black blocks are the actual regions, and the gray blocks are predicted exon regions.

The formulae are shown as follows

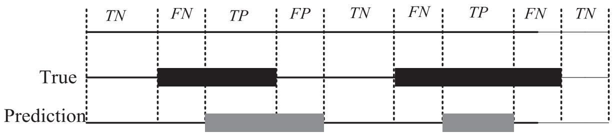

Figure 7 Evaluation of prediction accuracy at nucleotide level
The black blocks represent the actual coding regions, and the gray blocks represent the predicted exonic regions. TP: true positive; FP: false positive; TN: true negative; FN: false negative.

$$Sn = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{24a}$$

$$Sp = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \tag{24b}$$

$$P = 0.5 \times (Sn + Sp) \tag{24c}$$

To test the HMR195 data sets, correlation coefficient (CC) and approximate correlation (AC) are introduced which have been defined as

$$\mathrm{CC} = (\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP})/[(\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})]^{0.5}$$

$$\mathrm{ACP} = 0.25 \times [\mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) + \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) + \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) + \mathrm{TN}/(\mathrm{TN} + \mathrm{FN})] \tag{25b}$$

$$\mathrm{AC} = (\mathrm{ACP} - 0.5) \times 2 \tag{25c}$$

True positive (TP) is the number of coding nucleotides correctly predicted as coding regions. False negative (FN) is the number of coding nucleotides predicted as non-coding regions. True negative (TN) is the number of non-coding nucleotides correctly predicted as non-coding regions. False positive (FP) is the number of non-coding nucleotides predicted as coding regions [30]. In the evaluation performance, we compared the DFT method [8], anti-notch filter and multistage filter [16], the original SONF algorithm [23] and the modified wavelet technique [18]. The results of the comparison with the method reported previously [24] are shown in **Table 1**. Note that the listed data of DFT, anti-notch filter, multistage filter and the signal boosting based on DFT are obtained from previous study [29] using the same evaluation method. From the table, we observe that the percentages of prediction of the proposed model are obviously superior to other DSP methods. Maximum sensitivities of 0.721, 0.703, 0.673, 0.725 and 0.88 were obtained using the DFT technique, IIR anti-notch filter, multistage filter, signal boosting based method and modified Garbor-wavelet, respectively. Compared with the original SONF algorithm, the *Sn* is slightly higher while the *Sp* and precision *P* are enhanced significantly except for the time frequency method based on S-transform which can achieve an accuracy of 96%. In addition, the parameters value in this paper is slightly lower than the Markov-based model with high orders. However, it cannot outweigh the advantages of this proposed method. The SDG amplification can also effectively improve the accuracy of recognition.

Table 1 Evaluation performance (in %) of different methods for the *C. elegans* chromosome III

Gene prediction methods	*Sn*	*Sp*	*P*	References
[a]**Modified model**	**91.4**	**96.0**	**93.7**	This study
SONF model with $c = 0.5$	90.0	76.9	83.5	[23]
SONF with $c = 1$	90.0	51.7	70.8	[23]
DFT technique	72.1	39.4	89.7	Table 2 in [29]
IIR anti-notch filter	70.3	35.1	89.4	Table 2 in [29]
Multistage filter	67.3	26.6	88.5	Table 2 in [29]
Signal boosting based on DFT	72.5	47.1	91.1	Table 2 in [29]
Modified Garbor-wavelet	88.0	90.0	91.5	Table 1 in [18]
Time frequency method	**88.0**	**98.0**	**96.0**	Table 2 in [22]
Lengthen-shuffling FFT	78.8	79.9	79.3	Table 4 in [7]
Markov model $k = 1$	78.4	81.4	79.9	Table 4 in [7]
Markov model $k = 2$	85.4	94.5	89.9	Table 4 in [7]
Markov model $k = 4$	**91.9**	**95.6**	**93.8**	Table 4 in [7]
Markov model $k = 5$	**92.6**	**95.8**	**94.2**	Table 4 in [7]

Note: [a]Modified SONF model after Class II enhancement; data for Lengthen-shuffling FFT and Markov models with $k = 1$, 2, 4 and 5 are the best conditions where $P = (Sn + Sp)/2$ is used in evaluation. It is worthy noting that the values in bold face denote the superior recognitions. *Sn*, sensitivity; *Sp*, specificity; *P*, precision.

Also, the improved model has been used to test the HMR195 datasets in comparison with GeneMark, HMM, and FGENES programs [31]. **Table 2** lists the index parameters including *Sn*, *Sp*, *P*, *AC*, and *CC* analyzed using the evaluation scheme. The parameters in this table are the average values of every sequence in these datasets. From Table 2, the precision of HMMgen is 93.0%, while a precision of 90.7% is obtained in this paper. We should point out that the results for existing methods were obtained from the corresponding publications. Different methods have been tested on different gene datasets, the repetitive work of the existent methods were enormous and redundant. We simply use the results from the references. Therefore, two tables have been presented for two gene datasets.

The filter model based signal processing method require neither additional biological information or trained genomic datasets for prediction of coding regions, so it can be applied to analyze unknown and novel genomes. This paper focused on identification of long DNA sequences. From the simulations (see Tables and Figures), we find out that the proposed algorithm in this paper can effectively recognize the locations of coding regions. Although not as good as that of the high order Markov model, the results obtained using the proposed algorithm are acceptable. It is worth noting that the complexity of this filter model is lower than the high order Markov-based model (see the implementation program). In addition, the model in this paper shares some similarities with Kalman filter theory. However, the computing complexity of the modified algorithm is efficiently reduced without calculating the Jacoby matrix by using partial differential and modeling the status equations. Moreover, compared with the DFT and STDFT spectrum analysis, and time frequency methods [8–14,22], the window width *M* of the sliding window in this study does not require a multiple of 3 due to the power calculation of $S(N/3)$.

This paper aimed at investigating the validity and feasibility of the proposed model in genome analysis and prediction. To some extent, the Markov chain based model may be more effective in predicting coding regions by comparison with Ref. [4]. Markov model parameters were trained via a number of gene datasets, thus filter based methods cannot achieve its prediction ability and robustness, and the Markov chain model has widespread applications in many technical fields (*e.g.*, the practicable software on line). Although the prediction accuracy using the proposed model is slightly lower than that using the efficient

Table 2 Exon levels of HMR195 datasets from different gene finding programs

Programs of gene finding	*Sn* (%)	*Sp* (%)	*CC* (%)	*AC* (%)	*P* (%)
Filter model	**91.7**	**87.8**	**77.9**	**80.3**	**90.7**
GeneMark.HMM	87.0	89.0	83.0	84.0	88.0
HMMgene	**93.0**	**93.0**	**91.0**	**91.0**	**93.0**
FGENES	86.0	88.0	83.0	84.0	87.0
Genie	91.0	90.0	88.0	89.0	90.5
Morgan	75.0	74.0	69.0	70.0	74.5

Note: Data for GeneMark.HMM, HMMgene, FGENES, Genie, and Morgan were obtained from Table 1 [31]. For the HMR195 datasets, the HMMgene performs the best for recognition. It is worthy noting that the values in bold face denote the superior recognitions. *CC*, correlation coefficient; *AC*, approximate correlation.

Markov chain model with high orders, the filter model is superior to the prediction methods based on the frequency content (*e.g.*, DFT based techniques, IIR anti-notch filter and multistage filter) in terms of sensitivity, specificity and precision. We show that the improved filter model has reliable performance for exon prediction. We conducted the first independent comparative evaluation of the gene-finding algorithms available and designed a more convenient and simple algorithm for a broad approach to gene finding. Obtaining definitive accuracy seems to be an impossible task, since the performance of the programs is very sensitive to the datasets tested upon, as observed by many researchers. Not to mention that we have to assume that the actual coding exons were correctly annotated in the GenBank record under the "CDS" feature (annotated non-coding exons are not considered).

Conclusion

In this paper, a modified filter model is applied to detect protein-coding regions. To analyze gene sequences using signal processing theory, Z-curve representation is used to map DNA bases into digital sequences. Combined with the filter and the iteration algorithm, a sliding window is then applied for sampling gene data in order to analyze DNA sequences for predicting coding regions. We illustrated the potential use of the filter model to recognize a known DNA sequence. Our results strengthen its plausibility in detection of protein-coding regions. An advantage of the filter model is that it performs well without the limit of window width. In addition, the complex computation can be skipped without considering the Jacoby matrix. Another advantage is that the proposed filter model can achieve the identification of coding regions without any prior information about the DNA sequences.

To suppress the non-coding regions and enhance the SNR between coding regions and non-coding regions, a SDG amplification model with Class I and Class II amplification is carried out on the output of the filter model. Simulation results show that the coding regions can be clearly identified. An evaluation algorithm is then performed on the two models. Results show that the improved filter model in this paper is effective in predicting protein-coding regions. However, the SNR is supposed to be infinite in this new model and the gene datasets tested in this paper are from the gene library and thus can be thought as pure signal. Certain improvements and adjustments of the filter structure and more tests on noised gene data are desired for potential applications to the genome analysis including genome prediction and signal processing.

Authors' contributions

LZ conceived and carried out the project, and drafted the manuscript. FT supervised the study. SW participated in the design and coordination. All the authors have read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. CDJXS10160001) and the Central University Postgraduate' Science and Innovation Funds of China (Grant No. CDJXS12160005).

References

[1] Farber R, Lapedes A. Determination of eukaryotic protein coding regions using neural networks and information theory. J Mol Biol 1992;226:471–9.

[2] Kulp D, Haussler D, Reese MG, Eeckman FH. A generalized hidden markov model for the recognition of human genes in DNA. Proc Int Conf Intell Syst Mol Biol 1996;4:134–42.

[3] Hendeson J, Salzberg S, Fasman KH. Finding genes in DNA with a hidden Markov model. J Comput Biol 1997;4:127–41.

[4] Korf I. Gene finding in novel genomes. BMC Bioinformatics 2004;14:59.

[5] Ossadnik SM, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Peng CK, et al. Correlation approach to identify coding regions in DNA sequences. Biophys J 1994;67:64–70.

[6] Solovyev V, Salamov A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc Int Conf Intell Syst Mol Biol 1997;5:294–302.

[7] Gao F, Zhang CT. Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics 2004;20:673–81.

[8] Datta S, Asif A. A fast DFT based gene prediction algorithm for identification of protein coding regions. IEEE Int Conf Acoust Speech Signal Process 2005;5:653–6.

[9] Grandhi DG, Kumar CV. 2-Simplex mapping for identifying the protein coding regions in DNA. In: Proceedings of the IEEE Region Conference on TENCON, Tiapei, October 30–November 2, 2007, p. 1–3.

[10] Hota MK, Srivastava VK. DSP technique for gene and exon prediction taking complex indicator sequence. In: Proceedings of the IEEE region conference on TENCON, Hyderabad, 19–21 Nov, 2008, p. 1–3.

[11] Anastassiou D. Frequency-domain analysis of biomolecular sequence. Bioinformatics 2000;16:1073–81.

[12] Bergen SWA, Antoniou A. Application of parametric window functions to the STDFT method for gene prediction. In:IEEE pacific rim conference on communications,computers and signals processing (PACRIM), 2005, p.324–7

[13] Yan M, Lin ZS, Zhang CT. A new Fourier transform approach for protein coding measure based on the format of the Z curve. Bioinformatics 1998;14:685–90.

[14] Yin C, Yau SST. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J Theor Biol 2007;247:687–94.

[15] Ramachandran P, Lu WS, Antoniou A. Location of exons in DNA sequences using digital filters. IEEE Int Symp Circuit Syst 2009:2337–40.

[16] Vaidyanathan PP, Yoon BJ. Digital filters for gene prediction applications. In: Asilomar conference on signals,systems and computers, vol 1; 2002, p. 306–10.

[17] Guigó R. DNA composition, codon usage and exon prediction. Genetic Databases. New York: Academic; 1999, p. 53–80.

[18] Mena-Chalco J, Carrer H. Identification of protein coding regions using the modified Gabor-wavelet transform. IEEE/ACM Trans Comput Biol Bioinform 2008;5:198–207.

[19] Tuqan J, Rushdi A. A DSP approach for finding the codon bias in DNA sequences. IEEE J Select Top Signal Process 2008;2:343–56.

[20] Anastassiou D. Genomic signal processing. IEEE Signal Process Mag 2001;18:8–20.

[21] Wang SY, Tian FC, Liu X, Wang J. A novel representation approach to DNA sequence and its application. IEEE Signal Process Lett 2009;16:275–8.

[22] Sahu SS, Panda G. Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach. Genomics Proteomics Bioinformatics 2011;9:45–55.

[23] Kakumani R, Devabhaktuni V. Prediction of protein-coding regions in DNA sequences using a model-based approach. IEEE Int Symp Circuit Syst 2008:1918–21.

[24] Agarwal R, Plotkin EI, Swamy MNS. Statistically optimal null filter based on instantaneous matched processing. Circuit Syst Signal Process 2001;20:37–61.

[25] Yadav R, Agarwal R, Swamy MNS. A new improved model-based seizure detection using statistically optimal null filter. Conf Proc IEEE Eng Med Biol Soc 2009;2009:1318–22.

[26] Zhang R, Zhang CT. Z curve, an intuitive tool for visualizing and analyzing the DNA sequences. J Biomol Struct Dyn 1994;11:767–82.

[27] Turin GL. An introduction to digital matched filters. Proc IEEE 1976;64:1092–112.

[28] Fox TW, Carreira A. A digital signal processing method for gene prediction with improved noise suppression. EURASIP J Appl Signal Processing 2004;2004:108–14.

[29] Gunawan TS, Ambikairajah E, Epps J. A signal boosting technique for gene prediction. Proc IEEE ICICS 2007:1–4.

[30] Burest M, Guigo R. Evaluation of gene structure prediction programs. Genomics 1996;34:353–67.

[31] Rogic S, Mackworth AK, Ouellette FBF. Evaluation of Gene-Finding programs on mammalian sequences. Genome Res 2001;11:817–32.