Digital Signal Processing 46 (2015) 10–18

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A dynamic representation-based, *de novo* method for protein-coding region prediction and biological information detection

Sajid A. Marhon *, Stefan C. Kremer

School of Computer Science, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

ARTICLE INFO

Article history:
Available online 28 August 2015

Keywords:
Gene finding
Digital signal processing
Protein coding region
Protein non-coding regions
DNA dynamic representation scheme
Period-3 spectrum

ABSTRACT

In this paper, we propose a new method for the prediction of protein coding regions that is designed to detect novel genes that do not have known, close homologs. The proposed method uses a dynamic representation scheme to convert DNA sequences into a numerical form, and then it uses the nucleotide distribution variance to calculate the period-3 spectrum. The dynamic representation scheme assigns numerical pairs to the nucleotides to emphasize the effect of the nucleotides that have a stronger participation in the period-3 spectrum. The proposed method also uses the nucleotide distribution variance which has less computational cost than the Fourier transform to extract the period-3 spectrum. A post-processing of the period-3 spectrum signal is performed to smooth the signal, detect the period-3 spectrum peaks, and locate the boundaries of the protein-coding regions.

The analysis of the receiver operating characteristic (ROC) curves shows that the proposed method outperforms other Digital Signal Processing (DSP)-based methods. The analysis of the false positive peaks shows that these regions have a similarity with regions that have functional patterns in other DNA sequences. The method also highlights and explores the capabilities of techniques that perform better than homology-based techniques for *de novo* protein prediction. We believe that this is an area of research that has been underemphasized and deserves additional attention.

1. Introduction

Genomic sequences consist not only of classical protein coding gene regions but also of intergenic spacers as well. Much of this is probably junk DNA of little functional relevance, despite recent claims to the contrary. However, it has been revealed over the past 50 years that some of these intergenic regions are not only junk but they include many non-protein coding loci and other regions of biological importance. Some of these loci produce non-coding RNAs (ncRNAs) which are not translated, but have other roles in different processes. They affect the transcriptional regulation, chromosome structure, RNA processing and modification, gene splicing, and the stability and translation of mRNA [1,2]. ncRNAs include different classes such as the ribosomal RNA (rRNA), transfer RNA (tRNA), Piwi-interacting RNA (piRNA), micro RNA (miRNA), small interfacing RNA (siRNA), small nuclear RNA (snRNA), and so on [1]. There is a myriad of other types of non-coding regions as well, such as *cis*- and *trans*-regulatory elements, pseudogenes, repeat sequences, transposons, viral elements, and telomeres, to name a few. Typically, intergenic spacers constitute about 95% of a genomic strand, while protein-coding genes constitute only 2% of the human genome [3–5]. The low density of coding sequences in some eukaryotic genomes makes finding them particularly difficult.

Hidden Markov model (HMM)-based methods have proven very successful in predicting protein coding regions. Some of the most prominent HMM techniques are HMMgene [6], GenScan [7], Genie [8] and GeneMark.Hmm [9]. HMM techniques are trained on existing, labeled DNA sequences to differentiate between the compositional structures of protein-coding and non-coding regions. In this respect, they are very successful in detecting the boundaries of coding regions, and we consider the protein finding problem on the basis of homologs to essentially be a “solved” problem. The HMM performance, however, is coupled to the homology between novel DNA sequences to be labeled and those in the training datasets. Guigo [10] has stated that gene prediction techniques can be classified into model-dependent, such as HMM techniques, and model-independent, such as DSP-based techniques. The design of the model-dependent techniques is based on capturing the specific features of coding regions in DNA sequences by the process of estimating the model parameters. However, the design of model-independent techniques is based on capturing the universal features of coding regions with no need to estimate model

* Corresponding author.
E-mail addresses: smarhon@uoguelph.ca (S.A. Marhon), skremer@uoguelph.ca (S.C. Kremer).

http://dx.doi.org/10.1016/j.dsp.2015.08.007
1051-2004/

parameters for such models. Therefore, model-independent techniques can be more powerful when there is a lack of knowledge about the specific genomic structures of the species under consideration. In addition, a significant portion of the microbial diversity is still unexplored, and this is called microbial dark matter [11]. It is likely that, in exploring these microbial organisms, there will be new proteins discovered that are different than any previously observed. With the absence of any homologs for these proteins, homology-based gene prediction techniques are unlikely to be successful [11]. Just as in protein structure prediction, where there are two different approaches—homology-based methods (which perform well on proteins similar to those previously studied), and *de novo* methods (which are required when no convenient homologs exist)—we can say that in the field of gene prediction there is a need for a set of methods that are complementary to the homology-based methods, i.e. HMM approaches. In this paper, we show that, in the absence of homologs, DSP-based methods represent an effective alternative that can be used for gene prediction.

DSP has been applied to process DNA sequences for protein coding region prediction. The spectral analysis of DNA sequences shows that there is a strong period-3 spectrum peak in protein coding regions, whereas no such spectrum peak is observed in non-coding regions. One common tool of DSP that has been applied to extract the period-3 property is the short time Fourier transform (STFT). The STFT is a form of the discrete Fourier transform (DFT) that is applied to a segment of the signal. When the DFT is applied to a nucleotide sequence, the spectral analysis shows the prominence of the spectrum peak at frequency $N/3$, where N is the length of the segment to be processed [12–14]. The DSP-based gene prediction process includes four major steps: DNA sequence mapping, sequence windowing, period-3 property extraction, and signal thresholding. All of the four mentioned steps are important in the design of DSP-based gene prediction techniques and they affect the accuracy of the techniques [15].

Researchers have used different DNA representation schemes and post-processing to improve the discrimination between coding regions and non-coding regions in the analysis of the period-3 spectrum. Most of the techniques use an experimental, predefined threshold for this discrimination. Mena-Chalco et al. [16] used the modified Gabor wavelet transform to extract the period-3 spectrum. In the classification of the signal, they assumed an experimental percentage of the base pairs are coding regions, and this percentage is equivalent to the percentage of coding region density in DNA sequences. Jiang et al. [17] used the universal representation scheme to map DNA sequences and extracted the period-3 spectrum using the STFT. They used the mean of the period-3 signal as an experimental threshold to discriminate between coding and non-coding regions. Shakya et al. [18] proposed a post-processing algorithm to improve the gene prediction of transforms. The algorithm compares the period-3 signal of a DNA sequence with its period-3 suppressed version and the difference between them that is within a predefined threshold is classified as a non-coding region. In the mentioned literature and other DSP-based methods such as in [19,20], an experimental, predefined threshold is used to discriminate between coding and non-coding regions. Xu et al. [21] stated that organisms have different optimal thresholds that depend on the organisms' gene structure property. In other words, it is difficult to assume an experimental, predefined threshold value that can work properly for every organism. Therefore, this issue has motivated some researchers to propose dynamic thresholding instead of using a static, experimental threshold value to classify protein coding regions. Agrawal et al. [22] proposed an adaptive thresholding technique which dynamically determines a range of thresholds to classify the period-3 spectrum signal. The technique is based on fuzzy-logic rules that uses some statistical parameters such as the mean, standard deviation and maximum of the period-3 spectrum peaks to train the model. The fuzzy logic rules applied on the statistical properties are used to determine the threshold ranges. They help search for an optimal threshold value for the input signal.

Vaidyanathan and Yoon [23] proposed using the antinotch filter with a central frequency $\omega_0 = 2\pi/3$ to extract the period-3 spectrum. Variations of the antinotch filter have also recently been proposed by Hota and Srivastava [24]. Hota and Srivastava [24] used the antinotch filter with suppressing the conjugate frequency component. The conjugate frequency component contributes to the peak's strength in exons and introns causing inaccurate detection of coding regions [24]. Hota and Srivastava [24] also proposed using the antinotch filtering with harmonic frequency component suppression. Suppressing the conjugate frequency component and the harmonic frequency components can improve the prediction of protein coding regions [24]. Datta and Asif [25] used a fast DFT method to extract the period-3 component. They used a Bartlett window in the analysis instead of the rectangular window. The authors stated that the Bartlett window provides better attenuation of the extraneous peaks. They also determined an experimental threshold value to classify the coding regions and non-coding regions. Hsieh et al. [26] proposed the EXONSCAN technique which combines signal detection and coding region alignment. The technique first detects signals which are start codons, stop codons, splice acceptors and splice donors. The technique then compares the signals with another set of signals of another sequence by performing coding region alignment and checking the similarity. Therefore, this technique is homology-based and requires homologous information in order to predict coding regions.

Researchers have shown that the existence of the period-3 spectrum peak in coding regions could be due to the triplet nature of the nucleotides forming the codons of the protein amino acids [13,14]. Despite the trend of observing a period-3 spectrum peak in protein coding regions, the presence of a peak does not always indicate a protein coding region. While many researchers will simply dismiss this observation as noise and try to attenuate it in their techniques, we provide a more fulsome investigation of the phenomenon with some surprising results.

In this paper, we propose a new method for predicting protein coding regions based on the analysis of the period-3 spectrum. Our method uses the dynamic representation scheme for mapping DNA sequences into a numerical form. In addition, the nucleotide distribution variance tool, which is computationally less expensive than the DFT, is used to calculate the period-3 property. Furthermore, a new technique for thresholding the period-3 signal is proposed in this article. For simplicity, we will refer to the proposed method as the dynamic representation-based, *de novo* (DRdn) method. The remainder of this article is as follows. Section 2 presents our proposed method. Section 3 presents the simulations and the experimental results. Section 4 discusses the obtained results. Section 5 concludes the important points about the results and the proposed technique.

2. Methods description

In this paper, we propose the DRdn method to predict protein coding regions. The process of protein coding region prediction includes four important steps. The first step is the mapping of symbolic DNA sequences into a numerical form in order to process these sequences by DSP tools. The second step is the selection of the window length parameter used in the analysis by the STFT. The third step is the extraction of the period-3 spectrum using a DSP tool. Finally, the fourth step in the process is the thresholding of the signal to determine the coding and non-coding regions [15]. The MATLAB code of the proposed technique is available for academic use at http://smarhon.github.io/DRdn/.