2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

A New, Dynamic-Representation-Based Gene Finding Method With An Analysis of False Positive Peaks

Sajid A. Marhon and Stefan C. Kremer
School of Computer Science, University of Guelph
Guelph, Ontario, N1G 2W1, Canada
Email: {smarhon, skremer}@uoguelph.ca

***Abstract*—In this paper, we propose a new method for gene finding. The method uses a new dynamic representation scheme to map DNA sequences into a numerical form. The dynamic representation scheme assigns numerical pairs to the nucleotides based on their effectiveness in the period-3 spectrum. Nucleotides that have a stronger participation in the period-3 spectrum peaks are assigned numerical pairs that further enhance their participation. Another development that the proposed method introduces is the detection of the period-3 spectrum peaks to discriminate between protein coding and non-coding regions. In this paper, we also analyze the period-3 peaks that are predicted by the proposed method. We analyze the false positive peaks by scanning the stop codons in the possible reading frames. The work also analyzes the false positive peaks that are attached to true positive peaks. This analysis provides insights for future work that can be conducted to improve the prediction accuracy of spectrum-based techniques by studying the false positive peaks. In addition, it provides an insight about these false positive peaks that may have originated as transcribed sequences which, over time, acquired stop codons by mutation and lost their characteristic for transcription.**

***Keywords*—Gene finding, protein coding region, DNA sequences, DNA representation scheme, period-3 spectrum.**

I. Introduction

Gene finding is a fundamental step in bioinformatics that leads to studying and analyzing the structure and function of proteins. The prediction of genes or protein coding regions includes determining the lengths and boundaries of these regions in genomic sequences. Protein coding region prediction techniques are either model-dependent or model-independent techniques. Model-dependent techniques depend on training models and estimating some parameters to learn some specific statistical features of already annotated genes. Model-independent techniques capture the universal features of protein coding regions without training a model or estimating parameters [1]. Model-independent techniques are more suitable for *de novo* gene finding. These techniques are not trained on datasets of specific species and there are no parameters that are estimated based on the statistical features of the genomic sequences in these datasets. Since the design of model-independent (*de novo*) gene finding techniques is not based on particular datasets, these techniques are more suitable than model-dependent techniques for gene finding in novel sequences.

Digital Signal Processing (DSP)-based techniques are classified as model-independent techniques. These techniques analyze DNA sequences for the prediction of protein coding regions by extracting the "period-3" spectrum at frequency $f = 1/3$. The spectral analysis of protein coding regions shows a prominent spectrum peak at the frequency component mentioned above, whereas the analysis of non-coding regions does not show such a spectrum peak [2], [3], [4]. The prominence of this property in coding regions is due to the triplet nature of codons in these regions. One of the problems with the period-3 spectrum analysis is the false positive peaks in non-coding regions. Some non-coding regions have strong period-3 (false positive) peaks which obstruct the discrimination between coding and non-coding regions. We have previously conducted a study where we discovered that the false positives have close analogs that are coding regions and we suspect that these false positives might be deactivated pseudogenes [5].

Different representation schemes have been proposed and used to map DNA sequences into a numerical form in order to be processed by DSP tools. Most of these schemes use a static mapping in which nucleotides are assigned the same numerical values in the mapping process [6], [7], [8]. Mena-Chalco et al. [9], Tiwari et al. [10], and Yin and Yau [11] used the binary representation scheme which converts the four sequences corresponding to the four nucleotides to four binary indicator sequences. Jiang et al. [8] used a universal representation scheme which also assigns the same numerical value for each nucleotide. Shakya et al. [12] used the "paired numeric" representation proposed in [6] which maps a DNA sequence into a single numerical sequence. A comprehensive review about DNA representation schemes is published in [13].

Researchers have proposed different techniques to predict protein coding regions based on the spectral analysis of DNA sequences and used different classification methods to discriminate between coding and non-coding regions [13]. Mena-Chalco et al. [9] used the modified-Gabor wavelet transform to extract the period-3 spectrum. To classify the coding regions, they assumed an experimental percentage of the base pairs that have the highest period-3 spectrum values to be coding regions. This experimental percentage is equivalent to the percentile of coding regions in DNA sequences. Jiang et al. [8] used the mean of the extracted period-3 signal as a threshold value to classify the coding regions. Shakya et al. [12] proposed a post-processing algorithm to improve the accuracy of prediction. The algorithm calculates the spectrum with suppressing the period-3 component and compares it with the spectrum that in-

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.51

cludes the period-3 component. If the difference between them is within a predefined threshold value, the region is classified as a non-coding region. Tiwari et al. [10], and Yin and Yau [11] used an experimental threshold value in their methods to classify coding and non-coding regions. The techniques that use a predefined or experimental threshold value usually tune the threshold value to maximize the accuracy of prediction. We have noticed that there are some strong period-3 peaks that are not coding regions. We wonder why these regions might show up in spectral methods, but not be coding regions. In addition, a further study of these false positive peaks could lead to post-processing in spectrum-based gene prediction techniques that can help improve the prediction accuracy of these techniques.

In this paper, we are proposing a new Dynamic Representation-Based Gene Finding (DRGF) method. The method uses a new DNA representation scheme for the mapping of DNA sequences. In addition, it uses post-processing for the detection of the period-3 peaks instead of using a predefined threshold for the discrimination between protein coding and non-coding regions. The work also includes an analysis of the results of this technique by studying the false positive peaks in the prediction results. The remainder of this paper will be as follows. Section II presents the DNA spectrum analysis used for the prediction of protein coding regions. Section III presents the proposed DRGF method. Section IV presents the results and the experimental analysis. Section V discusses the results of the proposed method and the results of the analysis of the predicted period-3 peaks. Section VI concludes the important points and suggests future work.

II. DNA Spectrum Analysis for Gene Prediction

The Short-Time Fourier Transform (STFT) has been used to analyze DNA sequences for the prediction of protein coding regions. The STFT tool extracts the spectrum component at frequency $f = 1/3$. A genomic sequence (of length N) is processed based on the sliding window basis. Each time, a window of length L of the sequence is processed to calculate the spectrum at frequency $f = 1/3$ or the period-3 property. The calculated value of the period-3 property corresponds to the region of the base pair position located at the center of the processed window. In order to calculate the values of the period-3 spectrum for the whole sequence of N base pairs (bp), the sequence is zero-padded with $\left\lfloor \frac{L}{2} \right\rfloor$ at the beginning and with $\left\lceil \frac{L}{2} \right\rceil$ at the end. The computation of the period-3 spectrum at the base pair position i is as follows:

$$S_i = \sum_{\alpha} \frac{1}{L^2} \left| \sum_{k=0}^{L-1} u_{\alpha}(k+i)\, e^{-2\pi jk/3} \right|^2 , \quad i = 0, 1, \cdots, N-1 \quad (1)$$

where $\alpha \in \{A, C, G, T\}$, and u_{α} is the binary indicator sequence corresponding to nucleotide α.

III. The Dynamic-Representation-Based Gene Finding Method

In this paper, we propose a new gene finding method that uses a new, dynamic representation scheme for the mapping of DNA sequences into a numerical form. The proposed method also uses a post-processing step to detect the period-3 spectrum peaks instead of using a predefined threshold value for detecting protein coding regions.

A. Dynamic Representation Scheme

In the proposed dynamic representation scheme, four numerical pairs are used to convert DNA sequences into a numerical form. Using numerical pairs instead of single digits is intended to enhance the distribution variance among the three codon positions which is coupled to the period-3 property. The four pairs are selected to enhance the participation of the nucleotides in the period-3 property. Researchers have shown that the prominence of the period-3 property in protein coding regions is related to the biased usage of the nucleotides in the protein amino acids [10]. Moreover, it has been shown that the nucleotide distribution variance of the nucleotides in the three codon positions is the reason for the prominence of the period-3 spectrum peaks in protein coding regions [14], [15]. Yin and Yau [14] have shown that in a particular DNA sequence, one or two nucleotides have a stronger participation in the period-3 spectrum peaks than others.

We have selected four numerical pairs for the substitution of the four nucleotides. Two of the four pairs provide a larger variance than the other two. The four pairs that we have selected are as follow:

$$\begin{array}{ll} \text{Large Variance} & \left\{ \begin{array}{rr} 1, & 2 \\ -1, & -2 \end{array} \right. \\ \text{Small Variance} & \left\{ \begin{array}{rr} 1, & 0 \\ -1, & 0 \end{array} \right. \end{array} \quad (2)$$

In the proposed method, a DNA sequence is processed according the sliding window basis. In every window of the sequence, the effectiveness of the four nucleotides is evaluated. The nucleotides that have a strong participation in the period-3 spectrum are assigned numerical pairs that further enhance their participation in the period-3 spectrum. The evaluation of the effectiveness of each nucleotide is calculated as the nucleotide distribution variance of the nucleotide using the nucleotide distribution variance tool proven in [15]. The two nucleotides that have a larger nucleotide distribution variance are assigned the numerical pairs (1, 2) and (-1, -2) as these two pairs further enhance the participation of the two nucleotides in the period-3 property. The other two nucleotides are assigned the other two pairs which provide less variance among the three codon positions than the first two pairs. The evaluation and assignment are done in the processing of every window of the sequence, so the assignment of the numerical pairs is dynamic and a particular nucleotide could be assigned a different numerical pair in every window. For example, if we want to map the following segment of a DNA sequence:

ACAGTGACT

Assuming that nucleotides A and C have distribution variances larger than G and T. Nucleotides A and C will be assigned any of the pairs (1, 2) and (-1, -2), and nucleotides G and T will be assigned any of the pairs (1, 0) and (-1, 0). The absence of a nucleotide is substituted with the pair (0, 0). Therefore, the mapping of the four nucleotides for that segment will be as follows:

$$\begin{aligned} u_A = &\quad 1,2,0,0,1,2,0,0,0,0,0,0,1,2,0,0,0,0 \\ u_C = &\quad 0,0,-1,-2,0,0,0,0,0,0,0,0,0,0,-1,-2,0,0 \\ u_G = &\quad 0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0 \\ u_T = &\quad 0,0,0,0,0,0,0,0,-1,0,0,0,0,0,0,0,-1,0 \end{aligned} \tag{3}$$

B. Period-3 Spectrum Extraction

The nucleotide distribution variance tool, in [14], [15], has been used in the proposed method to evaluate the effectiveness of the nucleotide in order to be assigned a numerical pair by the dynamic representation scheme. In addition, it has been used to calculate the period-3 spectrum in a particular window of the sequence. A window of length L of the DNA sequence is shifted and the nucleotide distribution variance (period-3 spectrum) is calculated as follows:

$$\begin{aligned} S_\alpha(i) = F_{1i}^2 + F_{2i}^2 + F_{3i}^2 - \\ (F_{1i} \times F_{2i} + F_{1i} \times F_{3i} + F_{2i} \times F_{3i}) \end{aligned} \tag{4}$$

where F_{xi} is the sum of the numerical values at the reading frame x in the window starting at the i^{th} base pair position of the sequence u_α, and $i = 0,1,2,\cdots,N-1$. The window length will be equal to L when this formula is used to calculate the effectiveness of the nucleotide, and it will be equal to $2L$ when this formula is used to compute the period-3 spectrum as every base pair is substituted with a numerical pair. The period-3 spectrum is computed by summing $S_\alpha(i)$ over α.

C. Post-Processing

After computing the period-3 spectrum, post-processing is performed in this technique to detect the period-3 spectrum peaks. In most of the previous techniques, a predefined threshold value was used to classify the regions into coding and non-coding regions. The experimental threshold value has always been selected to maximize the prediction accuracy. Therefore, it is always selected in a way that overlooks the false positive peaks that appear in the prediction results without paying attention to analyze these peaks. Another disadvantage of using an experimental threshold value is that it is difficult to use an experimental threshold value that can work with every dataset or organism properly. Xu et al. [16] stated that organisms have different optimal threshold values and they depend on the gene structure properties of the organisms. Therefore, in the DRGF method, we use post-processing to detect the period-3 peaks instead of a threshold. The post-processing includes smoothing the period-3 signal to degrade the extraneous peaks which appear in non-coding regions. The period-3 spectrum peaks should survive the smoothing process. We have used a moving average smoothing filter as follows:

$$h_i = \sum_{k=-m}^{k=m} \omega_k f_{i+k}, \tag{5}$$

where $\omega_k = 1/(2m+1)$. m should be chosen properly so that the smoothing does not deteriorate the period-3 peaks in exon regions. Experimentally, we have found that the best value of m is about half the value of the window length parameter used in the STFT. After smoothing the signal, its numerical gradient is computed as follows:

$$\Delta = \frac{\partial P(x)}{\partial x}, \tag{6}$$

where P is the period-3 spectrum signal, x denotes the base pair position, and ∂ denotes the partial derivative symbol. The transition in the sign of the gradient signal, obtained by applying (6), from positive to negative (intersection with the x-axis) indicates the existence of the period-3 peaks at the place where the transitions take place. The location of the peak can be determined as the region from the positive peak to the negative peak of the corresponding transition.

IV. Methods and Results

The DRGF method has been applied to datasets to present its results and compare its performance with other methods. We have used a window size that is 351 bp. This window size is widely used by other DSP-based methods [10], [17], [18]. We have applied the DRGF method to three datasets to plot the Receiver Operational Characteristic (ROC) curves in order to compare its performance with other DSP-based techniques. The comparison has included DSP-based methods such as the Spectral Content Measure (SCM) [10], Modified-Gabor Wavelet Transform (MGWT) [9], Spectral Rotation Measure (SRM) [18], and the Optimized Spectral Content Measure (OSCM) [17]. The datasets that have been used in the analysis are the BG570, HMR195 and Asp67. The BG570 dataset includes 570 multiple-exon, single-gene sequences of vertebrates [19]. The HMR195 dataset includes 195 multiple-exon, single-gene sequences of human, mice and rats [20]. The Asp67 dataset includes 67 multiple-gene sequences of *Aspergillus fumigatus* [21]. The DRGF method has been applied to the three datasets. Figure 1 shows the ROC curves of the methods when they are applied to the BG570, HMR195 and Asp67 datasets collectively. The plots of ROC curves show that the DRGF method outperforms all other DSP-based methods.

In the analysis of the DRGF method, we have analyzed the period-3 peaks predicted by it. The BG570 and HMR195 datasets have been used in this analysis. We have scanned for stop codons in the regions where the predicted peaks are located. The scanning included the three possible coding frames. The peaks that have stop codons in all three possible reading frames are confirmed as false positive peaks. As the boundaries of the peaks are not accurate, we leave a number of base pairs – we call it the boundary threshold– at each side of the peak and scan the inside of the peaks to avoid stop codons that are located at the predicted boundaries of the peaks where stop codons would be expected. Figure 2 explains the search area for the stop codons after leaving n base pairs from each side, where n represents the boundary threshold. The peaks that are marked as false positives are considered non-coding regions. The sensitivity and specificity are then computed. The search for the stop codon has been done for different values of n such that $n = \{0,20,40,60,100,150,200,250,300,350,400,450\}$. The plots of the sensitivity versus the complement of the specificity (1-specificity) for each of the two datasets are shown in Figure 3. The point of the highest specificity (lowest 1-specificity) and lowest sensitivity occurs when the value of

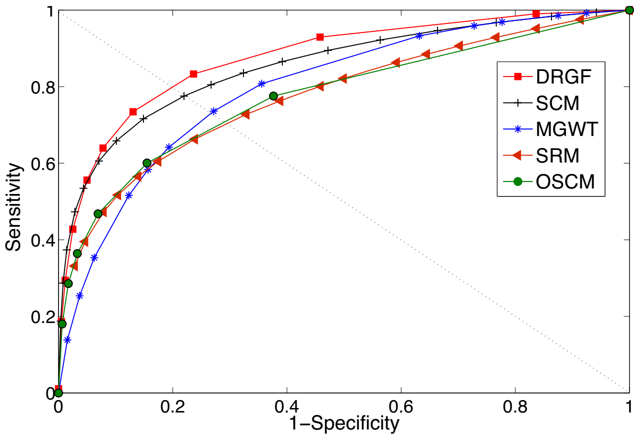

Fig. 1. The ROC curves of the DRGF, SCM, MGWT, SRM and OSCM methods when they are applied to the BG570, HMR195 and Asp67 datasets.

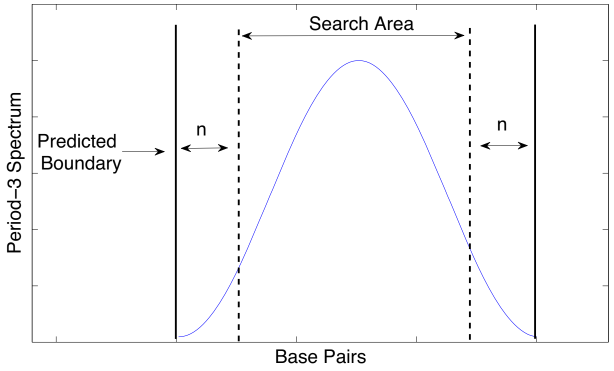

Fig. 2. The search area of the stop codon after neglecting n base pairs from each side of the predicted boundaries of the period-3 peak.

n is 0, while the point of the lowest specificity (highest 1-specificity) and highest sensitivity occurs when the value of n is 450.

We have conducted another experiment by setting the value of n based on the the length of the period-3 peak. The sensitivity and specificity have also been calculated using different values of n depending on the length of the predicted region. Table I presents the lengths of the period-3 peak and the corresponding values of n used in this experiment. We have filtered the peaks that have stop codons in the three possible reading frames. Therefore, some true positive peaks that are attached to or overlap with the filtered peaks are also filtered which decreases the value of sensitivity. Table II presents the results of this experiment for the BG570 and HMR195 datasets. The results of the BG570 dataset show that the number of the total predicted peaks is 2179, 1133 of them passed the filtering and 1046 peaks were filtered out. The value of the sensitivity was 0.42 and the value of specificity was 0.91. The average length of the filtered peaks is 585 bp and the average length of the unfiltered peaks is 355 bp. The low value of the sensitivity is because some of the filtered peaks are attached to or overlap with true positive peaks. Therefore, when the false positive peaks have been filtered out, the true positive peaks which are attached to or overlap with these false positive peaks have also been filtered and become false negatives. In order to investigate this issue thoroughly, we have conducted more analysis on the filtered peaks to compute the number of these peaks that are attached to or overlap with true positive peaks. We selected the peaks that overlap with at least 0.3 of their lengths from the beginning or from the end with a coding region. Table III presents the results of this experiment for the two datasets. The results of the BG570 dataset show that 542 peaks of the 1046 filtered peaks are attached to, or overlap with, true positive peaks. Therefore, these true positive peaks are forced to be false negatives in the filtering process which results in decreasing the sensitivity. When we exclude these peaks from the filtering process, the sensitivity goes up to 0.64 but the specificity decreases to 0.81 because the non-coding regions in these 466 peaks are now false positives.

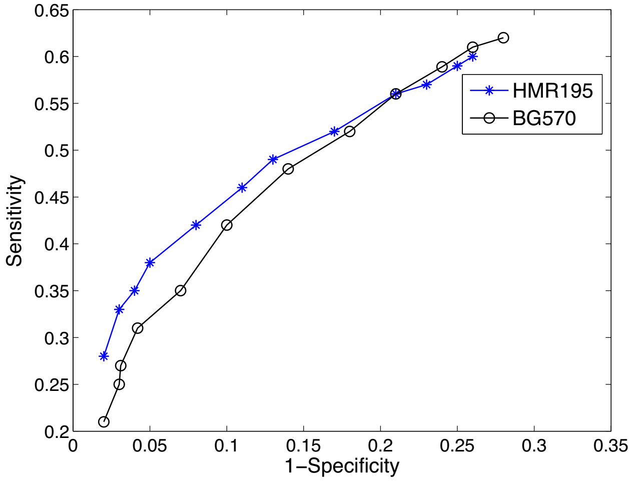

Fig. 3. The plots present the sensitivity versus the complement of the specificity (1-Specificity) of the two datasets when different values of base pairs n are left at the boundaries of the predicted peaks to determine the search area of the stop codons.

V. Discussions

In this article, we proposed a new gene finding technique based on the spectral analysis of DNA sequences. The technique used a new DNA representation scheme that dynamically assigned numerical pairs to the nucleotides. The mapping of DNA sequences into a numerical form is an important step in the processing of DNA sequences for gene finding [13]. In the proposed method, we considered the effectiveness of the nucleotides in the period-3 property. Some nucleotides have a strong participation in the period-3 property, while others have a weak participation. Therefore, we adopted enhancing the participation of the nucleotides that already have a strong participation by substituting them with numerical pairs that further enhance their participation. This discrimination between the nucleotides that are effective in the period-3 property and the others that are not is important to strengthen the period-3 peaks and make them well recognized from the background noise. This recognition of the period-3 peaks helped detect them easily for the purpose of classifying the protein coding regions.

TABLE I. THE ASSIGNMENT OF THE VALUES OF n BASED ON THE LENGTH OF THE PERIOD-3 PEAKS.

Peak Length	1-59	60-99	100-149	150-199	200-249	250-399	400-599	600-799	800-1199	>=1200
n	0	10	30	50	70	100	150	200	300	400

TABLE II. THE RESULTS OF THE TWO DATASETS WHEN THE VALUE OF n IS SET DEPENDING ON THE PEAK LENGTH.

Dataset	**Predicted Peaks**	**Passed Filtering**	**Filtered**	**Sensitivity**	**Specificity**
BG570	2179	1133	1046	0.42	0.91
HMR195	1014	547	467	0.46	0.91

TABLE III. THE RESULTS OF THE TWO DATASETS WHEN THE VALUE OF n IS SET DEPENDING ON THE PEAK LENGTH, AND THE PEAKS ATTACHED TO CODING REGIONS ARE EXEMPTED.

Dataset	**Peaks Attached to Coding Regions**	**Sensitivity**	**Specificity**
BG570	542	0.64	0.81
HMR195	168	0.62	0.85

The classification of DNA sequences into coding and non-coding regions is another important step in the process. In the DRGF method, we detected the period-3 peaks without using an experimental threshold value to discriminate between coding and non-coding regions. Our method can detect the period-3 peaks in a DNA sequence regardless of the variation in their strength. On the other hand, when a threshold value is used to classify coding and non-coding regions, weak period-3 peaks are always missed as false negatives. In our method, all the period-3 peaks can be detected because the gradients of the peaks are calculated and each peak can be detected regardless of the strength of that peak. In addition, when an experimental threshold value is set, it does not work well with all datasets. The optimal threshold value is different from one organism to another and it depends on the gene structure of the organism. Therefore, the detection of the period-3 peaks without using a threshold value is another advantage of the proposed method as it makes the discrimination between coding and non-coding regions more effective.

We have studied and analyzed the false positive peaks that appear in the prediction results of the DRGF method. These false positive peaks are recognized in the results of this method because the method does not use an experimental threshold value to discriminate between coding and non-coding regions. However, other spectral analysis-based methods set an experimental threshold value to avoid false positive peaks and maximize the prediction accuracy. Therefore, in this analysis, we have conducted experiments to study these false positive peaks and how they are related to the actual coding regions. Our analysis depends on the search of the stop codons in the DNA regions where the predicted peaks are located. The search for the stop codons includes the three possible reading frames as the datasets used in the analysis have genes in the forward strand and not in the complementary strand. As we are analyzing predicted peaks, the boundaries might not be accurate. Therefore, we have used a boundary threshold in which n base pairs are left from each side of the predicted peak and make the search in the remaining area of the peak region. We have considered the length of the predicted peaks when we set the boundary threshold. Peaks with longer regions are scanned with higher values of the boundary threshold. The experiment of scanning the stop codons in the predicted peaks has shown that almost half of the predicted peaks (1046 out of 2179) in the BG570 dataset have stop codons in the three possible reading frames. Similarly, in the HMR195 dataset, 476 of the 1014 predicted peaks have stop codons in the three reading frames. As we have forced the peaks that were filtered out to be "true negative", the more interesting than the absolute specificity and sensitivity are the changes in sensitivity and specificity as a result of stop-codon filtering.

Further analysis of the filtered peaks shows that more than half of them in the BG570 dataset are attached to, or overlap with, coding regions "true positive" peaks. When these filtered peaks are set as non-coding regions, the attached coding regions are also set as false negatives and that is why the sensitivity declined. Therefore, we have exempted the false positive peaks that are attached to coding regions form filtering. As a result, the sensitivity has increased but the specificity declined as shown in Table III. In this experiment, the analysis of the false positive peaks shows that some of these peaks have stop codons in the three possible reading frames. This means there is a possibility that they used to be coding regions but they lost their facility to be transcribed especially when the analysis shows that they are attached to coding regions. In addition, false positive peaks that do not have stop codons in one or more reading frame may also once have been coding regions but due to mutation the splicing has changed that made them lose the characteristic to be transcribed.

The analysis shows that some of the false positive peaks are attached to coding regions. The attachment of these false positive peaks to coding regions significantly affects the prediction accuracy of spectrum analysis-based gene prediction techniques even if a threshold value is tuned to discriminate between coding and non-coding regions. An important analysis or post-processing can be future work on how to separate the true positive peaks that are attached to false positive peaks and this step could improve the performance of such kind of techniques without degrading the sensitivity and the specificity.

VI. CONCLUSIONS

In this work, we have proposed a gene finding technique that used the spectral analysis of DNA sequences to predict protein coding regions. The proposed method adopted a new dynamic representation scheme that assigned numerical pairs to the nucleotides according to their effectiveness in the period-3 spectrum. Unlike other DSP-based methods, the proposed method used this dynamic representation scheme to enhance the strength of the period-3 spectrum peaks and provide better discrimination for these peaks from the background noise at non-coding regions. The other new thing that the proposed method has introduced is the detection of the period-3 spectrum peaks without using a predefined threshold value. Detecting the period-3 spectrum peaks for locating protein coding regions is more efficient than using a threshold value. Using a predefined threshold value does not work well with all datasets and with all organisms. Moreover, the variation in the strength of the period-3 peaks is a challenge when a threshold value is used to locate coding regions. However, all peaks in a sequence are detected in the proposed method regardless of their strength since the basis of detecting the peaks is based on computing the gradients of these peaks.

In this paper, we have conducted further analysis on the false positive peaks. The purpose of the study is to investigate the possibility of improving the performance of this technique and DSP-based techniques in general. Analyzing the false positive peaks can give a prospective post-processing technique that can improve the prediction accuracy of DSP-based techniques. Using a predefined threshold value in DSP-based gene prediction techniques could overlook some interesting insights about some false positive peaks that appear at non-coding regions in the prediction results. The analysis of the false positive peaks shows that it is possible to improve the prediction accuracy by separating the false positive peaks from the attached true positive peaks by conducting further analysis for a post-processing step as future work. This post-processing can improve the performance of prediction of DSP-based techniques. In addition, the existence of stop codons in these strong false positive peaks can give an insight that these regions have lost their characteristic to be transcribed as coding regions.

Acknowledgments

Dr. Stefan C. Kremer has been funded by the NSERC of Canada.

References

[1] R. Guigo, *DNA composition, codon usage and exon prediction: In Genetic Databases*. Academic Press, 1999, ch. 17, pp. 53–80.

[2] V. R. Chechetkin and A. Y. Turygin, "Size-dependence of three-periodicity and long-range correlations in DNA sequences," *Phys. Lett. A*, vol. 199, pp. 75–80, 1995.

[3] A. Tsonis, J. Elsner, and P. Tsonis, "Periodicity in DNA coding sequences: implications in gene evolution," *J. Theor. Biol.*, vol. 151, no. 3, pp. 323–331, 1981.

[4] J. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Res.*, vol. 10, no. 17, pp. 5303–5318, 1982.

[5] S. A. Marhon and S. C. Kremer, "A dynamic representation-based, *de novo* method for protein-coding region prediction and biological information detection," *In Preparation*, 2014.

[6] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction," in *Fifth IEEE International Workshop on Genomic Signal Processing and Statistics*, 2007, pp. 34–37.

[7] ——, "Signal processing in sequence analysis: advances in eukaryotic gene prediction," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 310–321, 2008.

[8] X. Jiang, D. Lavenier, and S. Yau, "Coding region prediction based on a universal DNA sequence representation method," *J. Comput. Biol.*, vol. 15, no. 10, pp. 1237–1256, 2008.

[9] J. Mena-Chalco, H. Carrer, Y. Zana, and R. Cesar Jr., "Identification of protein coding regions using the modified Gabor-wavelet transform," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 5, no. 2, pp. 198–207, 2008.

[10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Comput. Appl. Biosci.*, vol. 13, no. 3, pp. 263–270, 1997.

[11] C. Yin and S. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence," *J. Theor. Biol.*, vol. 247, no. 4, pp. 687–694, 2007.

[12] D. Shakya, R. Saxena, and S. Sharma, "Improved exon prediction with transforms by de-noising period-3 measure," *Digit. Signal Process.*, vol. 23, no. 2, pp. 499–505, 2013.

[13] S. A. Marhon and S. C. Kremer, "Gene prediction based on DNA spectral analysis: a literature review," *J. Comput. Biol.*, vol. 18, no. 4, pp. 639–676, 2011.

[14] C. Yin and S. Yau, "Fourier characteristics of coding sequences: origins and a non-Fourier approximation," *J. Comput. Biol.*, vol. 12, no. 9, pp. 1153–1165, November 2005.

[15] S. Marhon and S. C. Kremer, "Theoretical justification of computing the 3-base periodicity using nucleotide distribution variance." *BioSystems*, vol. 101, no. 3, pp. 185–186, 2010.

[16] S. Xu, N. Rao, X. Chen, G. Liu, and Y. Wang, "Analysis of threshold influence on the accuracy of gene-prediction methods based on power spectrum analysis," in *2010 IEEE 10th International Conference on Signal Processing (ICSP)*, 2010, pp. 1–4.

[17] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.

[18] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Res.*, vol. 13, no. 8, pp. 1930–1937, 2003.

[19] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, no. 3, pp. 353–367, 1996.

[20] S. Rogic, A. K. Mackworth, and B. F. Ouellette, "Evaluation of gene finding programs on mammalian sequences," *Genome Res.*, vol. 11, no. 5, pp. 817–832, May 2001.

[21] J. Mena-Chalco, "Available: http://www.vision.ime.usp.br/~jmena /mgwt/datasets/," July 2014.