Computational Identification of Short Initial Exons

Sayanthan Logeswaran[1], Eliathamby Ambikairajah[1], and Julien Epps[1,2]

[1] School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney 2052, Australia
[2] National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
sayanthan@ee.unsw.edu.au, ambi@ee.unsw.edu.au,
julien.epps@nicta.com.au

Abstract. Despite the existence of many gene prediction programs and their increasing accuracy over the last few years, the accurate identification of short exons remains a challenging problem. In this paper we concentrate on short initial exons and present a method to improve the detection of these short coding regions. The proposed algorithm is based on the Weight Array Method (WAM) and CpG islands. The algorithm was evaluated on a total of 158 sequences containing short initial exons, and achieves an accuracy of up to 73%. By comparison with GENSCAN, the proposed WAM-CpG Island algorithm reveals an improvement of up to 22%. Further, the WAM-CpG island approach can be employed to complement existing gene prediction packages to produce substantial improvements in the correct detection of short initial exons.

1 Introduction

Gene prediction in DNA sequences is a well-studied problem, and there are many gene prediction packages available [1],[2]. As more genomes are being sequenced, these gene prediction packages are becoming more sophisticated and more accurate [2], thus yielding more useful results to biological practitioners. All the current packages are able to predict most coding regions with a fairly high accuracy, but all exhibit weakness in their ability to detect short coding regions (fewer than 50 bases) [1],[3]. A very low accuracy is typically encountered when it comes to detecting these short exons, which leads to inaccurate gene prediction.

The root of the problem with most existing gene prediction packages is that protein coding regions (exons), are modeled using some kind of coding statistic. Coding statistics are measures indicative of protein coding regions, as certain statistical differences exist in coding and non-coding regions (introns) [11]. While these measures have proved to be effective in modeling most exons, they usually fail when dealing with short exons, as the coding length is simply too small to make an accurate distinction. Other techniques such as combining signal sensors and sequence similarity searches have made some improvements to the accuracy, but the detection of short coding regions remains poor.

In this paper, we propose a method for detecting short initial exons. We define a 'short exon' as an exon of length less than 50 bases, and we outline an algorithm that

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 42 – 48, 2006.

uses both the Weight Array Method and CpG islands to locate these exons. We compare the results of our algorithm with the popular gene prediction program GENSCAN [9], and also show how to combine our algorithm with a full gene prediction program in a complementary manner.

2 Methods

2.1 Datasets

We use three datasets in this paper, for parameter selection and evaluation of our proposed algorithm. These are the Burset/Guigo dataset of vertebrate sequences [7], the HMR195 dataset of mammalian sequences [1], and Reese/Kulp dataset of human sequences [8]. It is important to note that short exons occur infrequently in sequence databases (this fact contributes to the poor performance of existing gene prediction packages), so that relatively small quantities of data are available for experimenting on and benchmarking approaches to this problem.

2.2 Weight Array Method

We use a WAM to model the translation initiation signal and donor site of the initial exon. This model assigns a probability to an aligned sequence, in order to determine whether the sequence is a member of a particular class or not [4]. Given a sequence $s = (s_1,s_2,\ldots,s_N)$, where $s_i \in \{A, C, G, T\}$, around a potential start or donor site, we can distinguish whether the site is either a true or a decoy signal based on whether the probability

$$P(\mathbf{s}) = \sum_{i=1}^{N} w_i(s_i) + \sum_{i=1}^{N-1} \log \frac{P(s_i|s_{i+1})}{P(s_i)P(s_{i+1})}, \tag{1}$$

is greater than a pre-defined threshold.

In this paper we select this threshold by applying the WAM to all start and donor sites in the Reese/Kulp dataset. The start site was modeled using an $N = 19$ bp WAM from positions -12 to +6, with the start codon (ATG) occupying positions 0-2. The donor site was modeled using an N = 10 bp WAM from positions -4 to +5, with the GT dinucleotide occupying positions 0-1. The threshold was then selected based on the true positive to false positive ratio.

2.3 CpG Islands

The precise definition of CpG islands is somewhat arbitrary, but generally speaking CpG islands are regions in a DNA sequence that are rich in the dinucleotide CpG. These islands are very useful in gene prediction as they are known to be associated with the start of genes [5]. We follow the definition employed in [6], and define a CpG island as a DNA sequence of length greater than 200 bp, with a high C+G content and a frequency of CpG dinucleotides close to the expected value, where the ratio is calculated as

$$R = \frac{Observed\ CpG}{Expected\ CpG} = \frac{Num_{CpG}\ M}{Num_C\ Num_G}, \quad (2)$$

where Num_X is the number of occurrences of $X \in \{A, C, G, T\}$, and M is the length of the sequence window. In this paper we will define the overall CpG island score S_{CpG} as the average of the C+G content and the above ratio

$$S_{CpG} = \tfrac{1}{2}\left(Num_C + Num_G + R\right). \quad (3)$$

This score is obtained by moving a sliding window of length M = 200 bases, from –500 bases upstream to +500 bases downstream of a potential translation initiation site. The maximum of the resulting CpG island scores is defined as the CpG island score for that sequence.

2.4 Proposed Algorithm

The aim of the proposed algorithm is to accurately identify short initial exons. Since the length of these exons is too small to employ any coding statistic, we look for features outside the exon to identify these regions. The most obvious signals to look for are the start site and donor site. Thus, given a new DNA sequence, we first search for all possible start sites. If the start site is within the predetermined threshold, we then search for all possible donor sites up to 50 bases downstream of the successful start site. If this donor site is also within the predetermined donor threshold, we mark the two as a potential exon. Thus we end up with a set of potential initial exons. Now, if we combine the start site and donor site scores of each exon, and sort the potential exons according to this combined score, we find empirically that the true exon is usually in the top 3. By way of demonstration, we extracted all sequences containing short initial exons from all three datasets mentioned in section 2.1 and applied the above method, with the results shown in Table 1.

Over 75% of the true initial exons were contained in the three-best scoring for the databases considered. Thus, the next step is to extract the true exon from the top three. We achieve this by examining the other signals around the initial exon. As the initial exon is located around the promoter region there are many other signals to look for, such as CAP sites, TATA boxes and CpG islands. We chose CpG islands, since they have been shown to be the most dominant signal in computational promoter detection [12].

Table 1. Percentage of true exons in top three

Dataset	No. Seq	% Initial Exons in Top 3
Burset/Guigo	97	86.60
Reese/Kulp	53	79.25
HMR195	31	58.06
Overall	158	76.58

Thus, the top three exons are selected and the CpG island score S_{CpG} around their start site is calculated. Using this CpG island score and the combined WAM scores, the true exon is selected. Fusion of the WAM and CpG island scores is achieved by calculating the difference,γ, between the top two combined WAM scores. If $\gamma$is large, the highest score is chosen to be that of the true exon. If $\gamma$is small, the true exon is chosen to be that with the highest scoring CpG island score from the top three exons. From empirical tests, we found the optimal value for γ to be 1.5. A summary of the algorithm is given as follows:

Algorithm 1. WAM-CpG island algorithm

Step 1: *Search for all possible short initial exons.*
Step 2: *Sort them according to the combined WAM score.*
Step 3: *Select top 3 and calculate their CpG island score.*
Step 4: *If the difference between top 2 WAM scores is large choose the top WAM score as the true exon*
Step 5: *If the difference between top 2 WAM scores is small choose the top CpG island score exon from the top 3 exons as the true one.*

3 Results and Discussion

3.1 Evaluation Measures

We evaluated the accuracy of the proposed WAM-CpG Island algorithm using the conventional exon Sensitivity (*Sn*) and Specificity (*Sp*) measures defined in [7]. Sensitivity is the proportion of actual exons that are correctly predicted, and is defined as

$$Sn = \frac{Number\ of\ \ Correct\ Exons}{Number\ of\ \ Actual\ Exons}, \tag{4}$$

Specificity is the proportion of predicted exons that are correctly predicted, and is defined as

$$Sp = \frac{Number\ of\ \ Correct\ Exons}{Number\ of\ \ Predicted\ Exons}. \tag{5}$$

3.2 Results

We tested the algorithm on the three different datasets described in section 2.1. Again, we extracted all sequences containing initial short exons from these datasets, computed our algorithm and calculated the accuracy measures as defined above. The results are shown in Table 2.

Table 2. WAM-CpG island algorithm on sequences containing short initial exons

Dataset	No. of Seq	*Sn*	*Sp*
Burset/Guigo	97	72.16	73.68
Reese/Kulp	53	69.81	69.81
HMR195	31	51.61	51.61
Overall	158	68.99	69.87

By way of comparison with GENSCAN, we tested the same sequences on GENSCAN, and obtained the results shown in Table 3.

Table 3. GENSCAN accuracy on sequences containing short initial exons

Dataset	No. of Seq	*Sn*	*Sp*
Burset/Guigo	97	58.76	82.61
Reese/Kulp	53	47.17	64.10
HMR195	31	29.03	69.23
Overall	158	48.73	75.49

A clear improvement in the detection of short initial exons was observed using the proposed WAM-CpG Island algorithm. GENSCAN appears to completely miss or partially predict most short initial exons. Again, the reason for the lack of accuracy in GENSCAN and indeed most gene prediction programs is that coding statistics play a large role in their exon prediction models. These statistics are modeled on training sets that do not favor short exons [1]. As short exons are not as common as other exons, current prediction programs have no features to specifically deal with them and thus these programs fail to pick them up. Hence the success of the proposed algorithm lies in the fact that it specifically looks for short exons and nothing else, and thus if a short initial exon does exist there is a high probability that the algorithm will detect it.

3.3 Combining Algorithms

The sensitivity and specificity measurements in Table 2 and Table 3 demonstrate how well the proposed algorithm works on sequences that contain short initial exons. In order to obtain the full benefit of our algorithm for practical purposes, it needs to be applied to a full dataset. Hence, in this section we show how to combine the proposed WAM-CpG Island algorithm with a full gene prediction package, taking GENSCAN, and the HMR195 dataset as our example. The proposed method is illustrated in Figure 1.

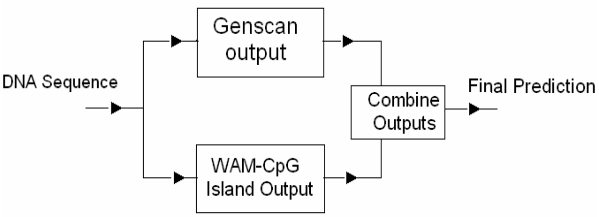

Fig. 1. GENSCAN/WAM-CpG island combining algorithm

We then compared the accuracy of the predicted initial exons of GENSCAN with and without combining the WAM-CpG Island algorithm.

Thus, 195 mammalian sequences from the HMR195 dataset were processed using GENSCAN. From the 195 sequences, only 152 sequences contain initial exons, of which 31 are short exons. The results of testing the 152 sequences on GENSCAN are shown in Table 4.

Table 4. GENSCAN accuracy on full dataset

	No. of Seq	Correct Init. Exons
Full Dataset	152	56.58%
Short Exons	31	29.03%

We then combined our algorithm with the GENSCAN output as follows. Firstly, we note that GENSCAN defines any exon it predicts with probability < 0.5 to be weak and unreliable, so that it should be 'treated with caution' [9], and also recall the earlier observation that GENSCAN completely misses or partially predicts most short initial exons. Thus the following criteria are used to alter the initial exon predictions obtained by GENSCAN:

1. If no initial exon is predicted and the WAM-CpG Island prediction is before and in phase with the succeeding exons, choose the WAM-CpG Island prediction.
2. If no initial exon is predicted and the predicted first internal exon is weak, replace the predicted first internal exon with the WAM-CpG Island prediction if it is before and in phase with the succeeding exons.
3. If the GENSCAN predicted initial exon is weak, replace the predicted initial exon with the WAM-CpG Island prediction if it is before and in phase with the succeeding exons.

Using the above criteria, the WAM-CpG Island algorithm was applied to the GENSCAN output to obtain the results shown in Table 5.

Table 5. Combined accuracy on full dataset

	No. of Seq	Correct Init. Exons
Full Dataset	152	60.53%
Short Exons	31	48.39%

Thus, from Table 5, there is a clear improvement in the detection of short initial exons when combining our algorithm with GENSCAN, which as a result also improves the overall accuracy of initial exons for the dataset. By incorporating a short exon algorithm into a full gene prediction package, the detection of exons and genes can be significantly improved.

Although GENSCAN has specifically been used as our example to demonstrate the use of the WAM-CpG island algorithm, it should be straightforward to similarly incorporate it with any gene prediction package, and achieve similar results.

4 Conclusion

This paper has introduced a new algorithm, based on the Weight Array Method and CpG islands, for the detection of short initial exons. The proposed algorithm was compared with GENSCAN, and a clear improvement in the detection accuracy for such short exons was observed. A method for combining the proposed WAM-CpG

Island algorithm with a full gene prediction package in a complementary manner was then given, and an improvement in the overall accuracy of initial exon prediction was demonstrated. Future work aims at further strengthening this algorithm, by using more robust signal sensor models via support vector machines [10] in place of the WAM, and incorporating more promoter features such as TATA boxes to reduce the false positives. Other work may include employing a similar approach for the detection of short terminal exons.

Acknowledgement

This research is fully supported by the University of New South Wales, Australia, Goldstar Award, 2005 for genomic signal processing research.

References

1. Rogic, S., Mackworth, A. K. and Ouellette, F.B.F.: Evaluation of Gene-Finding Programs on Mammalian Sequences, Genome Research, vol. 11 (2001), pp. 817-832.
2. Brent, M.R., and Guigo, R.: Recent advances in gene structure prediction,. Curr. Opin. Struct. Biol., 14(3) (2004), pp. 264-272.
3. Mathe, C., Sagot, M.F., Schiex, T. and Rouze, P.: Current methods of gene prediction, their strengths and weaknesses, Nucl. Acids Res., vol. 30, no. 19 (2002), pp. 4103-4117.
4. Zhang, M.Q. and Marr, T.G.: A weight array method for splicing signal analysis, CABIOS, vol. 9, no. 5 (1993), p. 499-509.
5. Bird, A.: CpG islands as gene markers in the vertebrate nucleus, Trends Genet., vol. 3 (1987), pp. 342−347.
6. Gardiner-Garden, M., and Frommer, M.: CpG islands in vertebrate genomes, J. Mol. Biol., vol. 196 (1987), pp. 261−282.
7. Burset, M., and Guigo, R.: Evaluation of gene structure prediction programs, Genomics, vol. 34 (1996), pp. 353-357.
8. http://www.fruitfly.org/sequence/human-datasets.html
9. Burge, C. and Karlin, S.: Prediction of complete gene structure in human genomic DNA, J. Mol. Biol. 268: 78–94 (1997).
10. Zien, A., *et al*: Engineering support vector machines that recognize translation initiation sites, Bioinformatics, vol. 16, no. 9 (2000), pp. 799-807.
11. Guigo, R.: DNA composition, codon usage and exon prediction, www.pdg.cnb.uam.es /cursos/FVi2001/GenomAna/GeneIdentification/SearchContent/main.html (2000).
12. Hannenhalli, S., and Levy, S.: Promoter prediction in the human genome, Bioinformatics, vol 17, Suppl. 1 (2001), pp. s90-s96.