CHAPTER 4
DIGITAL SIGNAL PROCESSING METHOD FOR ACCEPTOR SPLICE SITE PREDICTION

Most of the eukaryotic genes consist of the exonic regions (protein coding regions) separated by intronic regions. These intronic regions are removed by splicing (conversion from DNA to mRNA) and only exonic regions join together to form a contiguous gene for the protein synthesis. The splice sites (acceptor splice sites and donor splice sites) play an important role in the accurate detection of the protein coding regions in the DNA sequences as they mark boundaries between exons and introns. In this chapter, a method for the detection of acceptor splice sites has been proposed. The acceptor splice site (3’ end of intron) is the exon-intron border and the last two nucleotides of intron are considered as consensus dinucleotide “AG” as shown in Fig. 4.1. But the detection of acceptor splice sites is very difficult because of occurrences of the dinucleotide “AG” at the locations other than the acceptor sites in a whole gene sequence. [Akhtar et al. (2006)].

![Fig.4.1 Acceptor splices sites](image)

For the identification of the splice sites statistical methods that have been reported in the literature are weight matrix method (WMM) [Stadadn et al. (1984)], weight array method (WAM) [Zhang et al. (1993)], and windowed weight array method (WWAM) [Burge et al. (1998)]. As these methods are based on probabilities of occurrence of the nucleotides they are data dependent methods and also require training prior to making the predictions. Signal processing methods have been used to identify the exonic regions by exploiting the period-3 behavior [Tiwari et al.(1997), Anastassio et. al.(2001), Guan et al.(2004), Rao et al.(2004),]
Ambikairajah et al. (2005), Epps et al. (2005), and Akhtar et al. (2005a, 2008), Sahu et al. (2010), Shakya et al. (2011), Shakya et al. (2013a, 2013). Since the exonic region begins from the last two nucleotides of the intron which are considered as consensus dinucleotide “AG” or acceptor site in the 3’ end direction, the presence or absence of the periodicity-3 property in this region can be utilized for the recognition of the candidate acceptor sites. Therefore, any period-3 detection method can be employed for the identification of the acceptor splice sites in the DNA. Akhtar et al. (2006) employed a windowed discrete Fourier transform (WDFT) method because WDFT based method does not require the training of the genomics data before its use. The performance of this method however is not better than the other reported statistical methods. But by combining the DFT method with WAM method it has been reported that it improves the detection performance [Akhtar et al. (2006)]. In this chapter the adaptive STFT and PCA method developed in Chapter 3 has been used for the detection of the acceptor splice sites. The performance of the proposed method has been compared with the WDFT method.

4.1 Period-3 Spectra Using Windowed Discrete Fourier Transform

The WDFT method has been reported for the identification of the acceptor sites in the DNA sequences [Akhtar et al. (2006)]. WDFT is one of the traditional method for exon prediction. To compute WDFT of the DNA sequence, Voss mapping scheme is used to convert the nucleotide sequence in to numeric sequence and then DFT is applied to get the power spectrum with sliding window approach [Marhan et al. (2011)]. The N-point DFT of a numeric sequence \( x(n) \) is defined as:

\[
x(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{-j2\pi nk/N}
\]

where, \( k = 0, \ldots, N - 1 \), \( w(n) \) is the rectangular window of fixed length. Using (4.1), the period-3 power spectrum of the sequence is:

\[
S(k) = |x(k)|^2
\]

The power spectrum \( S(k) \) has large peaks at the frequency bin \( k = N/3 \) for the nucleotide positions where the exons are present. Optimal window length of “69” has been reported for better resolution and sharpness of the peaks. The period-3 spectrum obtained using WDFT for M60858 sequence is shown in Fig.4.2.
4.2 Period-3 Spectrum using Adaptive STFT and PCA

The method to obtain period-3 spectrum using adaptive STFT and PCA has been described in chapter 3. The period-3 spectrum using Adaptive STFT is noisy than the period-3 spectrum of the WDFT and is shown in Fig.4.3.

**Fig.4.2** Period-3 spectrum using WDFT

**Fig.4.3**: Period-3 spectrum using Adaptive STFT
**4.3 Score for Acceptor Sites**

To calculate the score for true acceptor splices sites, a very small window of length 70bps has been selected here because candidate acceptor site region is very small. The location of the window is the central position of the consensus dianucleotide “AG” and score for the true acceptor site has been calculated using following relation -

\[
Score = \log_2 \left( \frac{\sum_{i}^{lc} p_{3}}{\sum_{i}^{unc} p_{3}} \right) \tag{4.3}
\]

Where, \( p_{3} \) is the period-3 spectrum, \( uc \) and \( lc \) are the upper and lower limits of the coding region respectively and \( unc \) and \( lnc \) are the upper and lower limits of the noncoding regions. The parameter used in the score calculation has been described in Fig.4.4.

![Fig.4.4: Calculation of score at each splice site](image)

The scores for the WDFT based method and AST method has been plotted for the sequence M60858 in Fig4.5 and Fig.4.6 respectively.
The score calculated using adaptive STFT with PCA is relatively less noisy than WDFT method for the consensus dinucleotide ‘AG.'
4.4 Data Set and Evaluation Method

Following datasets and evaluation parameter have been used for the performance analysis of proposed method.

4.4.1 Data set

The GENSCAN [Burge et al. (1998)] dataset has been used for the experiment and one of its sequences with acc. no. M60858 has been used to demonstrate the proposed method for the acceptor site prediction. The GENSCAN dataset contains 386 acceptor splice sites. We labeled the nucleotide positions for A & G at ‘-2’ and ‘-1’ respectively. The last intron nucleotide positions have been labeled as -70, -69, and upto -1 for true acceptor splice site of the sequence, whereas the first exon nucleotides positions have been labeled as +1, +2, and upto +70. Here, a assumption has been made that the minimum length of the exon is 70 bps.

4.4.2 Performance evaluation criteria

Following The performance of the proposed method has been evaluated using following criteria’s-

- Firstly, the number of false positives have been calculated at different levels of percentage sensitivity [Shakya et al. (2011)]. A threshold score ‘Th’ is the minimum score at a particular level of percentage sensitivity ‘s’ for which percentage sensitivity of the true acceptor sites have scores greater than ‘Th’. The sensitivity is the ratio of the True positive (TP) and True positive (TP) plus False negatives (FN).

- An ROC curve shows the relation between ‘TP’ and ‘FP’ at each level of the threshold and for the best discrimination ROC curves should be more closure to ‘1’ or move upwards. If it is more closure to diagonal or moves down then the discrimination will be poor. AUC has also been calculated and it is associated with accuracy of the applied method. If AUC is more than accuracy will be high. [Youden et al.(1950)].

4.5 Results and Discussion

Using evaluation criteria we also obtained the TP, FP, FN, TN values and % sensitivity for the example sequence with acc. no. M60858 and plotted the graph between FPs and % sensitivity for WDFT and adaptive STFT with PCA in Fig.4.7. From Fig.4.7 it is observed that the adaptive STFT with PCA detects less number of FPs than WDFT at sensitivities.
Fig. 4.7: Variation of FPs with % sensitivity for M60858 sequence

Fig. 4.8: ROC curves for the sequence M60858

In Fig. 4.8 ROC curves for the two methods have been plotted. Again the proposed method provides a better performance than WDFT. The AUC values and FPs at different sensitivity levels are given in Table 4.1 for both the methods under study.
Table 4.1 Performance parameters for the M60858 sequence

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>% Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>FP</td>
</tr>
<tr>
<td>WDFT</td>
<td>0.6497</td>
<td>74</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.7257</td>
<td>28</td>
</tr>
</tbody>
</table>

The adaptive STFT based method and WDFT based methods have been also tested on the GENSCAN test set of 64 sequences. The numbers of false positives have been calculated at different level of percentage sensitivity and these are plotted in Fig.4.9.

![Variation of False positives with % sensitivity for GENSCAN test set](image)

**Fig.4.9** Variation of False positives with % sensitivity for GENSCAN test set
Fig. 4.10 ROC Curves for GENSCAN test Set

Fig. 4.9 shows the plots of the number of FPs at different levels of percentage sensitivity for these two methods using GENSCAN test set. The proposed method gives lower number of FPs. For better discrimination the FPs should be small at different sensitivity levels. Fig. 4.10 shows the ROC curve for WDFT and proposed method using GENSCAN test set. ROC curves for the proposed method moves up and remains above the other curve, depicting better discrimination power in the detection of the acceptor splice sites. Area under the ROC curves and the number of FPs at % of sensitivity have also been calculated and summarized in the Table 4.2 for these two methods for the GENSCAN dataset.

Table 4.2 Performance parameters for the GENSCAN test set

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>% Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP</td>
</tr>
<tr>
<td>WDFT</td>
<td>0.6772</td>
<td>1632</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.7257</td>
<td>1777</td>
</tr>
</tbody>
</table>
4.6 Summary
In this work, a digital signal processing based method has been proposed for the identification of the acceptor splice sites in the DNA and its performance has been compared with the reported WDFT method. The proposed method provides an improvement in the acceptor splice sites detection performance over the WDFT. The method is model independent. By combining it with statistical methods its detection performance can be further enhanced.