PREDICTION OF TRANSCRIPTIONAL START SITES OF GENES USING ASYMPTOTIC LOCAL APPROACH

C. W. Chan and W. K. Yeung

Department of Mechanical Engineering
The University of Hong Kong, Pokfulam Road, Hong Kong, China

Abstract: A popular approach to detect the Transcriptional Start Sites (*TSS*s) of genes that are *CpG* sensitive is based on the *CpG* islands. For these genes, their *TSS*s is characterized by sudden increases in the *CpG*s in the DNA sequence. In this paper, a novel gene prediction method is proposed that transforms the problem of detecting the TSSs to that of detecting a change in the mean of a stochastic process using the asymptotic local approach. Features of the *CpG* islands, such as the cyclic nature, are used to reduce the false detection rate. The proposed method is applied successfully to identify all the genes in the rabbit alpha-like gene cluster, and in a section of the human chromosome 22, 73% of the confirmed genes are predicted. Comparison with the Dragon Gene Start Finder is also made showing that the proposed method has a higher sensitivity.

Keywords: Transcriptional Start Sites, gene detection, C_pG islands, asymptotical local approach

1. INTRODUCTION

The outcome of the sequencing of DNA is a string of four different bases: adenine (*A*), thymine (*T*), cytosine (*C*), and guanine (*G*). After sequencing, the next important task is to extract the hidden meaning behind these data, by first finding the genes, then determining how their expression is regulated and the functions of the proteins they encode. These results would be very useful for analyzing the association between gene mutations and diseases, and for discovering measures to cure the diseases.

There are two common approaches to identify genes. The first approach is based on comparison with known genes, whilst the second one is to identify special features in the *DNA* sequence signifying the possible existence of genes. In the first approach, picking the gene sequence for comparison and selecting the thresholds in the comparison can be difficult and time consuming. Further, these methods suffer from the main drawback that it is virtually impossible to identify new genes, as comparisons are made only with known genes.

In the second approach, criteria are proposed to discriminate between the protein-coding regions and the non-coding regions. Since these techniques do not rely solely on the knowledge of existing genes, they are therefore more general. Several criteria are proposed: differences in codon usage (Staden and McLachlan, 1982), hexamer counts (Fickett, 1996), entropy measures (Almagor, 1985). Statistical analysis methods are also developed based on neural networks (Snyder and Stormo, 1995), or Markov models (Burge and Karlin, 1998), or a combination of these techniques (Bajic *et al.*, 2002).

Another popular criterion to find genes is the detection of the Transcriptional Start Site (*TSS*). Among the signals for identifying the *TSS*, the proximity to *CpG* islands is recognized to be one of the most important ones (Hannenhalli and Levy, 2001), as they overlap the promoter and extend about 1,000 base pairs (*bp*s) downstream into the *TSS* of a gene. The *CpG* islands are characterized by sudden increases in the counts of *C* followed immediately by *G* in the *DNA* sequence, and they are unmethylated regions of the genome that are associated with the 5' ends of most house-keeping genes and many

regulated genes (Birds, 1986). The *CG* binucleotides are often referred to as the *CpG* to reflect the phospodiester bond that connects the two nucleotides. The best of existing techniques can only achieve in eukaryotic genomes a sensitivity of about 60%, and a specificity of 50% (Westhead *et al*., 2002).

As sudden changes in the count of *CpG*s that characterizes the *CpG* islands are similar to faults in engineering systems, the detection of *TSS*s can be formulated as a fault detection problem, and well developed fault detection techniques for engineering systems can then be applied to detect genes in the *DNA* sequence. Asymptotic local approach, which transforms the fault diagnosis problem into one that detects statistical changes in the mean of a random variable, has been applied successfully to detect faults in engineering systems (Wang and Chan, 2002), and is extended to detect the *CpG* islands in the *DNA* sequence in this paper. Two case studies involving the rabbit alpha-like globin gene cluster and a section of the human chromosome 22 are presented. In the case of human chromosome 22, 73% of the confirmed genes can be predicted using the proposed method.

The organization of the paper is as follows. In Section 2, a brief review of the *DNA* sequence, the *CpG* islands, and the asymptotic local approach is present-ed. The detection of the *TSS*s in the *DNA* sequence based on the asymptotic local approach is derived in Section 3, and its performance is illustrated using the rabbit alpha-like globin gene cluster and a section of the human chromosome 22, as presented in Section 4. For the human chromosome 22, a comparison with the Dragon Gene Start Finder method is also made.

2. PRELIMINARIES

2.1 The CpG Islands

A genome is a complete set of instructions for making an organism containing the master blueprint for all cellular structures and activities for the lifetime of the cell or organism (Krane and Raymer, 2002). The *DNA* molecule in human genome, similar to other higher organisms, is a two-strand wrapping around each other that resembles a twisted ladder. A single strand of *DNA* is generated by chaining together nucleotides via a phosphodiester bond. In this bond, the phosphate molecule of a nucleotide is attached to the hydroxyl group of the next nucleotide. The 5' end of a nucleotide is a phosphate group attached to the 5' carbon of the pentose sugar. The 3' end is the nucleotide that is free for appending to the next nucleotide. The pairing of the bases follows certain strict rules, e.g., adenine will pair only with thymine (the *A*-*T* pair) and cytosine with guanine (the *C*-*G* pair). The ordering of the bases along the sugar-phosphate backbone is referred to as the *DNA* sequence, and this sequence contains the exact genetic instructions for a particular organism with its own unique traits.

The human genome contains approximately 3 billion *bp*s, organized into 24 distinct and physically separate microscopic units called chromosomes. All genes are arranged linearly along the chromosomes. The nucleus of most human cells contains two sets of chromosomes inherited from the parents. Each set of chromosomes contains 23 single chromosomes composing of 22 autosomes and an *X* or *Y* sex chromosome. Under normal condition, a female has a pair of *X* chromosomes and a male has an *X* and *Y* pair. In the chromosomes, it contains roughly equal parts of protein and *DNA*, which contains an average of 150 million bases. The *DNA* molecules are among the largest molecules now known.

As an example, the number of *CpG*s in the following section of a *DNA* sequence is 6.

GATCATCATCGAATGGAGTTGAATGGAATTA
TCAAAGAATGGAATCCAGTGGTATCATCATC
AAATGGAACCGAATGGAATCATCAAATGGAC
TCAAATGGAATCATTGAATAGATTCGAATGG
AATCATCATCGAATGAAATCGAATGGAAAAA
TTGAATGGACTCGAATGGAACCATCATTGAA
TGGAAACCAAAGGA

2.2 The Local Asymptotic Approach

Consider the nonlinear system,

$$y(t) = f(y(t-1), \cdots, y(t-n)) + e(t) \qquad (1)$$

where $y(t)$ is the output, $e(t)$ is a zero mean white noise with a variance of σ^2, and n is the maximum lags in the output. Define $x(t) = [y(t - 1) \ldots y(t - n)]^T$, where $x \in R^n$. Assuming system (1) is approximated by a nonlinear model, $\hat{f}(x(t))$, obtained by minimizing $\|y(t) - \hat{y}(t)\|$ for $x \in R^n$. Let

$$\hat{y}(t) = \hat{f}(x(t)) \qquad (2)$$

From (1) and (2), the modeling error, $\varepsilon(t)$, is

$$\varepsilon(t) = y(t) - \hat{y}(t) = f(x(t)) - \hat{f}(x(t)) + e(t) \qquad (3)$$

For a system with component or actuator fault, and sensor fault, the output of the system becomes (Wang and Chan, 2002),

$$y(t) = f_f(x(t)) + \Delta y(t) + e(t) \qquad (4)$$

where $f_f(x(t))$ is the nonlinear dynamics of the system after a component or actuator fault has been developed, and $\Delta y(t)$ arises from the sensor fault. There are a number of approaches to approximate the nonlinear system (1), including neural networks and neurofuzzy networks (Wang and Chan, 2002). From (4), the component or actuator fault can be formulated as changes in the weights of the network, whilst sensor faults, a change in the output. Since the model given by (2) is trained from data collected before any faults have developed, the modeling error $\varepsilon(t)$ obtained from this model is:

$$\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t) \\ &= f_f(x(t)) - \hat{f}(x(t)) + \Delta y(t) + e(t) \end{aligned} \qquad (5)$$

As some of the *TSS*s are characterized by a sudden increases in *CpG*s, similar to the case of a sensor fault, methods developed for detecting sensor faults

can therefore be extended to detect the *TSS*s. In this case, $f_f(x(t)) \approx \hat{f}(x(t))$, and (5) reduces to,

$$\varepsilon(t) = \Delta y(t) + e(t) \qquad (6)$$

Since $e(t)$ is assumed to be zero mean, the problem of detecting the *TSS*s can be formulated as one that detects the change in the mean of $\varepsilon(t)$, similar to the case of detecting sensor faults. Clearly, the ability to detect $\Delta y(t)$ from $\varepsilon(t)$ depends on the amplitude ratio of the noise $e(t)$ to $\Delta y(t)$. The detection of $\Delta y(t)$ is now reduced to the detecting of a change in the mean of a Gaussian process, and the asymptotic local approach is shown to be effective to detect this class of changes (Wang and Chan, 2002). Consider

$$\begin{aligned} D_M(t) &= \frac{1}{\sqrt{M}} \sum_{i=0}^{M-1} \varepsilon(t-i) \\ &= \frac{1}{\sqrt{M}} \sum_{i=0}^{M-1} (\Delta y(k-i) + e(k-i)) \end{aligned} \qquad (7)$$

Since $e(k)$ is assumed to be a Gaussian process with zero mean, hence $D_M(t)$ is also a Gaussian process with zero mean, if $\Delta y(t)$ is identically zero. However, if $\Delta y(t)$ is non-zero, the mean of $D_M(t)$ is clearly non-zero. Consequently, the detection of $\Delta y(t)$ is now reduced to that of detecting a change in the mean of $D_M(t)$. Define $S_M(t)$,

$$S_M(t) = \frac{D_M{}^2(t)}{\sigma^2(t)} \qquad (8)$$

where $\sigma^2(t)$ is the variance of $e(t)$ at time t. It is well known that $S_M(t)$ is χ^2-distributed with 1 degree of freedom (Wang and Chan, 2002). Therefore, the changes in the mean of $D_M(t)$ can now be detected by χ^2-test for a given confidence limit.

3. GENE PREDICTION BASED ON THE ASYMPTOTIC LOCAL APPROACH

Following the discussion in Section 2.1, a time-series, denoted by $N_{cg}(t)$, is formed from the *DNA* sequence by first dividing it into sections with a size of n, and then aggregating the number of *CpG*s in each section. The argument t of $N_{cg}(t)$ is the section number, and is also the order of the aggregated data in the time-series. The choice of n is important as it affects the effectiveness of the method to predict the *TSS*s. Since the occurrence of *CpG*s can be random, $N_{cg}(t)$ is generally a noisy time series. If n is chosen to be too small, there will be many zeros in $N_{cg}(t)$, and if n is chosen to be too large, not only the number of data for the statistical test is reduced, special features in $N_{cg}(t)$ may also be smoothed out. In both cases, it would be difficult to detect the *TSS*s.

As shown in Fig. 1(a) in Example 1, $N_{cg}(t)$ is quite erratic, and approximating it by a time-varying nonlinear model is difficult. A simple, yet effective approach is to consider it to be a time-varying stochastic process with a time-varying mean, which can be estimated by a moving window with a fixed size, as follows,

$$\hat{f}(t) = \frac{1}{n_m} \sum_{i=1}^{n_m} N_{cg}(t-i+1) \qquad (9)$$

where n_m is the size of the moving window. The estimated modeling error $\hat{\varepsilon}(t)$ is given by,

$$\hat{\varepsilon}(t) = N_{cg}(t) - \hat{f}(t) \qquad (10)$$

For a suitably chosen n_m, $\hat{\varepsilon}(t)$ is usually reasonably random with zero mean, except at the *TSS*s, where there is a sudden increase in *CpG*s. Consequently, the detection of *CpG* islands is reduced to detecting a change in the mean of $\hat{\varepsilon}(t)$ from zero. The asymptotic local approach described in Section 2.2 is used to detect this change. First, $D_M(t)$ is computed from (7) for a given M. Then $S_M(t)$ is computed from (8), and the χ^2 test is performed at a given confidence level. Similarly, the estimated variance of $\hat{\varepsilon}(t)$, denoted by $\hat{\sigma}^2(t)$, is also computed by a moving window with a fixed size, as follows,

$$\hat{\sigma}^2(t) = \frac{1}{n_v} \sum_{i=1}^{n_v} (\varepsilon^2(t-i)) \qquad (11)$$

where n_v is the window size. Let λ be the threshold obtained from the χ^2-table for a given confidence level. Then a positive prediction of a *TSS* is made, if $S_M(t)$ exceed this confidence level, and a gene is likely to be found in the vicinity of the predicted *TSS*. The sensitivity of this technique clearly depends on the choice of the confidence level. A large confidence level, i.e., close to 1, makes the test more stringent, and consequently, less false predicted *TSS*s will be produced. In contrast, a smaller confidence level increases the sensitivity of the method, but more predicted *TSS*s will be obtained, and the number of false predictions will increase. It should be noted that the strategy for choosing the confidence level in the prediction of the *TSS*s is different from that fault detection. In engineering systems, false fault detections may be costly. However, false prediction of genes is not too serious a problem, as it only increases the experimentation costs, which is often relatively small. However, the cost of missed predict-ion of genes is much higher, as important genes may not be found quickly. For these reasons, the strategy for choosing the confidence level is to increase the probability of predicting genes, but without unduly increasing the number of false predictions.

3.1 Guidelines for Choosing the Design Parameters

The design variables to be chosen by the users are n, n_m, n_v and M, which are the window sizes for computing $N_{cg}(t)$, the moving mean $\hat{f}(t)$, the moving variance $\hat{\sigma}^2(t)$ and the statistics $D_M(t)$, and also the confidence level in the χ^2 test. The following discussion concentrates mainly on the choice of the design parameters for predicting *TSS*s in the human chromosomes to be used in case study 2 in this paper. A common choice of n is 200 (Gardiner-Garden and Frommer, 1987). The window size n_m for calculating the moving mean can be chosen to be 20. As the data in $N_{cg}(t)$ are obtained by aggregating 200*bp*s, setting n_m to 20 implies 4000 *bp*s are used to calculate $\hat{f}(t)$. A characteristic of the data in $N_{cg}(t)$ is that it often

contains sequences of consecutive zeros, especially if n is smaller than 200. Therefore, it is important to select a suitable n_v to compute the estimated variance $\hat{\sigma}^2(t)$. Too small a n_v can lead to a small $\hat{\sigma}^2(t)$. Consequently, $S_M(t)$ computed by (8) becomes statistically significant in the χ^2 test, leading to excessive false predictions of *TSS*s. The window size for calculating the moving variance $\hat{\sigma}^2(t)$ is set to 15, i.e., 3,000 *bp*s are used to estimate $\hat{\sigma}^2(t)$. Since a large number of data is required at the start of the computation, and to avoid the problem of false prediction arising from unreliable and small estimate of $\hat{\sigma}^2(t)$, a lower bound for $\hat{\sigma}^2(t)$ is proposed here. If the section of data in $N_{cg}(t)$ for computing $\hat{\sigma}^2(t)$ is sufficiently random, it is reasonable to assume at least half of the data in the moving window n_v contain a single count of C+G. After correcting for the mean of 0.5, the estimated variance becomes

$$\overline{\sigma}^2 = \frac{n_v \times 0.5^2}{n_v} = 0.25 \qquad (12)$$

Consequently, if $\hat{\sigma}^2(t)$ computed from (11) is less than $\overline{\sigma}^2$, then $\hat{\sigma}^2(t)$ is replaced by $\overline{\sigma}^2$. In the χ^2 test, M is set to 5 and the confidence level to 99.9%. If higher sensitivity is required, the confidence level can be lowered to 99%. It should be noted that these suggested values are guidelines only, and should be fine tuned for specific applications.

3.2 Prediction of the TSSs

From the χ^2 test, a prediction of a *TSS* is made, if $S_M(t)$ exceeds λ, the threshold obtained from the χ^2-table for a given confidence level. To reduce false predictions, the following techniques are used. Since there is a higher count of the *CpG*s at the *TSS*s, $\hat{f}(t)$ is cyclic, as shown in Fig. 1(b) in Example 1 with one gene in each cycle. If there are more than one statistically significant $S_M(t)$ in a cycle, then a *TSS* is predicted at the position of the largest $S_M(t)$. Also, predicted *TSS*s that are near the trough of the moving mean $\hat{f}(t)$ are also ignored, since as explained earlier, the *TTS*s corresponds to regions with large number of *CpG*s.

From (8), $S_M(t)$ is a function of the square of $\hat{\varepsilon}(t)$, and hence it can be statistically significant for negative $\hat{\varepsilon}(t)$. However, as discussed previously, the predicted *TSS* is expected to occur when there is a sudden increase, not a sudden decrease in *CpG*s. Therefore, statistically significant $S_M(t)$ with negative ε(t) are also ignored.

3.3 Predicting Genes in the Reverse Direction

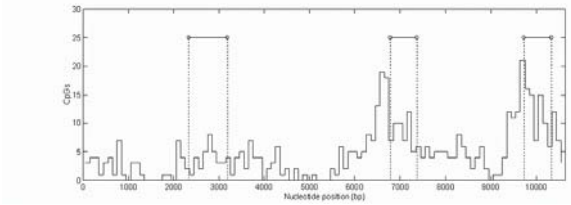
The same procedure can be applied to predict genes in the reverse direction, i.e., the 3'end of the *DNA* sequence. Since *C* and *G* are complementary, a *CpG* dinucleotide pair in the direct strand implies that there is also a *CpG* dinucleotide pair in the reverse strand. Consequently, a time series similar to $N_{cg}(t)$ with the data arranged in the reverse order is obtained first, before applying the procedure described in Section 3.3 to detect the *TSS*s. Since a prediction in either the forward or reverse direction often indicates a *TSS* either in the 5' end or in the 3', it seems sensible to try to find the genes in both directions whenever a prediction of the *TSS* is made.

4. CASE STUDIES

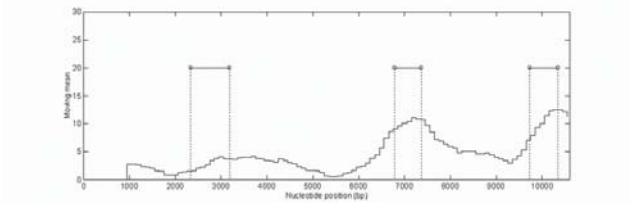
Case Study 1–Rabbit alpha-like globin gene cluster

The rabbit alpha-like globin gene cluster is well known to be *CpG* islands rich (Hardison, *et al*., 1991). Since it is a rather short sequence, a different set of parameters from that described in Section 3.3 are used. They are: n = 100, n_m = 10, n_v = 5, and M = 3. The aggregated *CpG*s time series, $N_{cg}(t)$, is plotted in Fig. 1(a), together with three confirmed genes marked by a solid line at the top of the graph. The moving average $\hat{f}(t)$ shown in Fig. 1(b) has three cycles, same number as the number of genes in the sequence. The residuals $\hat{\varepsilon}(t)$ computed by (10) are plotted in Fig. 1(c), and $S_M(t)$ computed by (8) in Fig. 1(d). At a confidence level of 99.5%, the threshold obtained from the χ^2-table for 1 degree of freedom is: λ = 7.88, marked by a horizontal dotted line. From Fig. 1(d), four statistically significant $S_M(t)$ are observed for t at the nucleotide positions of 2150, 5650, 6650 and 9750. From discussion in Section 3.2, $S_M(t)$ at t = 5650 is ignored, as there is another more significant one at 6750 in the same cycle. Hence, all three confirmed genes are correctly predicted. Note that although the *CpG*s islands for the first confirmed gene are not as obvious as that for the other two, the proposed method is able to predict it successfully.

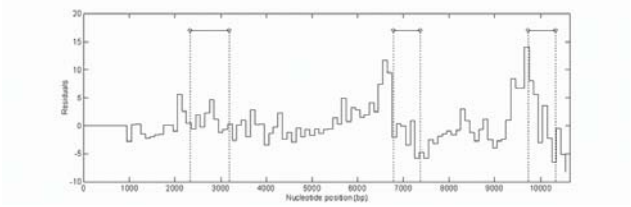
(a) Aggregated *CpG*s: $N_{cg}(t)$



(b) Moving mean $\hat{f}(t)$



(c) Residuals $\hat{\varepsilon}(t)$



(d) $S_M(t)$

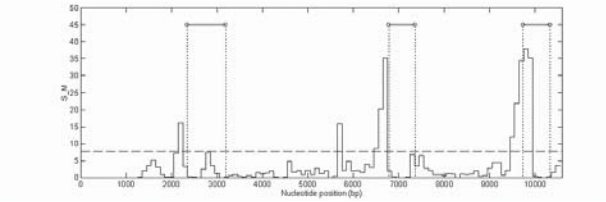


Fig. 1 Aggregated *CpG*s, moving mean, residuals and $S_M(t)$ for rabbit alpha-like globin gene cluster

Case Study 2 Human chromosome 22

A section of the human chromosome 22, labeled as NT_011519.10 is considered here. It consists of 3,661,581 *bp*s and is available at the National Center for Biotechnology Information (NCBI) website: http://www.ncbi.nlm.nih.gov/genome/guide/human/. The genes are categorized into 5 groups: (1) confirmed genes, (2) prediction + Expressed Sequenced Tags (EST) evidence, (3) EST evidence only, (4) prediction only, and (5) interim Locus ID. The order of the groups indicates the likelihood of genes that can be eventually found. For example, it is more likely to find a gene from predictions in Group (2) than that in Group (5).

From Section 3.1, the design parameters are: $n = 200$, $n_m = 20$, $n_v = 15$, and $M = 5$. In the following analysis, a true positive prediction refers to the case that an annotated gene is found within 2,000 *bp*s from the predicted position, and a false positive, if no annotated gene is found. If a known annotated gene fails to be predicted by the method, then a false negative is said to be made (Bajic, 2000).

The results of the prediction of these 5 groups of genes in both forward and reverse directions by the proposed method are presented in Table 1. The performance of the proposed method is assessed by the sensitivity and specificity. The former criterion is the ratio of the number of true positives to the number of annotated genes. The closer the sensitivity is to 1, the higher is the ability of the method to predict genes. The latter criterion is the ratio of the number of true positives to the total number of predictions. The higher is this ratio, the more likely that predictions obtained by the method is a true prediction. Ideally, it is desirable that both indices are as high as possible. However, from experience, for a method devised from one approach, it is likely that an increase in one of these indices often leads to a decrease in the other. Therefore, methods devised using several approaches may be necessary to achieve better performance in both indices. In gene prediction, sensitivity is more important than specificity, as low specificity only increases the experimentation costs, whilst a low sensitivity implies that the method is unable to perform the task of predicting genes.

Table 1 Prediction of genes by the proposed method

	Categories	*True Positive*	*False Negative*	*True total*	*False Positive*
1	Confirmed	40	15	55	
2	Predicted + EST evidence	9	11	20	
3	EST evidence only	2	4	6	
4	Predicted only	0	2	2	
5	Interim Locus ID	10	15	25	
	Overall	61	47	108	65

From Table 1, the sensitivity and specificity are computed, and are shown in Table 2. The results for Groups 1 and 2 are more important, as Group 1 consists of confirmed genes, and genes are more likely to be found eventually in Group 2 than the remaining groups. A relative high sensitivity of 0.65 is obtained for genes in both Groups 1 and 2. The specificity, as discussed earlier, is only 0.39.

Table 2 Sensitivity and specificity of the proposed method

Categories	*Sensitivity*	*Specificity*
1	0.73	0.32
1+2	0.65	0.39
1+2+3	0.63	0.40
1+2+3+4	0.61	0.40
1+2+3+4+5	0.56	0.48

The proposed method is now compared with the Dragon Gene Start Finder (*DGSF*) (Bajic and Seah, 2003). It is a computer program that uses three methods to predict genes: (i) the Dragon Promoter Finder system that makes use of promoter sensor, exon sensor and intron sensor, (ii) the estimation of the presence of the *CpG* islands, and (iii) sensor fusion methods that combines information obtained from (i) and (ii) using data preprocessing techniques and artificial neural networks.

The predictions of the genes by the *DGSF* are summarized in Table 3, and the sensitivity and specificity are shown in Table 4. Comparing Tables 2 and 4, the sensitivity of the proposed method are higher than that of the *DGSF*, though the specificity of the proposed method is lower. This is probably because several methods are incorporated into the *DGSF* that reduce the false predictions, illustrating the advantage of combining several techniques to improve the specificity of gene prediction methods. However, as the *DGSF* also uses information on the *CpG* islands, and yet the sensitivity of the *DGSF* is inferior to the proposed method that utilizes only the *CpG* islands indicates that the information from the *CpG* islands have not been fully utilized in the *DGSF*.

Table 3 Prediction of genes by the *DGSF*

Categories	*True Positive*	*False Negative*	*True total*	*False Positive*
1	39	16	55	
2	3	17	20	
3	1	5	6	
4	0	2	2	
5	7	18	25	
Overall	50	58	108	33

Table 4 Sensitivity and specificity of the *DGSF*

Categories	*Sensitivity*	*Specificity*
1	0.71	0.47
1+2	0.56	0.51
1+2+3	0.53	0.52
1+2+3+4	0.52	0.52
1+2+3+4+5	0.46	0.60

5. CONCLUSIONS

As *TSS*s are characterized by sudden increases in the *CpG*s, similar to sensor faults in engineering systems, statistical tests for sensor faults based on the asymptotic local approach is extended to detect these sites in this paper. The proposed method is simple to use, as it only involves choosing design parameters such as the window sizes for aggregating the *CpG* data, and for calculating the moving mean, the moving variance, and $S_M(t)$ in the χ^2 test. Other features, such as the cyclic nature of the moving mean are utilized to remove false predictions, and hence to increase the specificity of the method. The proposed method is applied to the rabbit alpha-like globin gene cluster and a section of the human chromosome 22. It is shown that all the genes in the former sequence have been successfully identified. In the human chromosome 22, the proposed method is able to predict 73% of the confirmed genes in the chromosome. The proposed method is compared with the Dragon Gene Start Finder. It is shown that the proposed method has a higher sensitivity, but a lower specificity. This is not too serious a drawback, as the penalty for not being able to predict possible genes is much higher than the increase in experimentation costs. As information of the *CpG* islands is better utilized, the proposed method is a better platform for incorporating other approaches to develop better method to predict genes.

REFERENCES

Almagor, H. (1985). Nucleotide distribution and the recognition of coding regions in DNA sequences: An information theory approach, *J. Theor. Biol.*, **117**, pp. 127-136.

Bajic, V.B. (2000). Comparing the success of different prediction software in sequence analysis: A review, *Briefings in Bioinformatics*, **1(3)**, pp. 214-228.

Bajic, V. B., Chong, A., Seah, S. H. and Brusic, V., 2002, An Intelligent System for Vertebrate Promoter Recognition, *IEEE Intelligent Systems*, 17(4), 64-70.

Bajic, V.B. and S.H. Seah (2003). Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes, *Nucleic Acids Research*, **31(13)**, pp. 3560-3563.

Bird A.P. (1986). CpG-rich islands and the function of DNA methylation, *Nature*, **321**, pp. 209-213.

Burge, C. and S. Karlin (1997). Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, pp. 78-94.

Burge, C.B. and S. Karlin (1998). Finding the genes in genomic DNA, *Curr. Opin. Struct. Biol.*, **8**, pp. 346-354.

Fickett, J. W., 1996, The gene identification problem: An overview for developers, *Comput. Chem.*, 20, 103-118.

Gardiner-Garden, M. and M. Frommer (1987). CpG islands in vertebrate genomes, *J. Mol. Biol.*, **196**, pp. 261-282.

Hannenhalli, S. and S. Levy (2001). Promoter Prediction in the Human Genome, *Bioinformatics*, 9th International Conference on Intelligent Systems for Molecular Biology, Denmark, pp. S90-S96.

Hardison, R., D. Krane, D. Vandenbergh, J-F. Cheng, J. Mansberger, J. Taddie, S. Schwartz, X. Huang and W. Miller (1991). Sequence and comparative analysis of rabbit α-like globin gene cluster reveals a rapid mode of evolution in a G+C-rich region of mammalian genomes, *J. Mol. Biol.*, **222**, pp. 233-249.

Krane, D.E. and M.L., Raymer (2002). *Fundamental Concepts of Bioinformatics*, Benjamin Cummings.

Snyder, E.E. and G.D. Stormo (1995). Identification of protein regions in genomic DNA, *J. Mol. Biol.*, **258**, pp. 1-18.

Staden, R. and A.D. McLachlan (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences, *Nucleic Acids Res.*, **10**, pp. 141-156.

Wang, Y. and C.W. Chan (2002). Asymptotic local approach in fault detection with faults modeled by neurofuzzy networks, *Proceedings of IFAC World Congress '02*, Barcelona, Spain, July 21-26.

Westhead, D.R., J.H. Parish and R.M. Twyman (2002). *Bioinformatics*, Oxford: BIOS.