Chandrashekar.Jatoth, Rupesh Mahajan / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.247-254

An*G*-HPR: Analysis of *n*-Gram based human Promoter Recognition

Chandrashekar.Jatoth[1], Rupesh Mahajan[2]
Department of Information Technology, Pad.Dr.DY.P.I.E.T, Pune, India

Abstract

We describe a promoter recognition method named An-HPR to locate eukaryotic promoter regions and predict transcription start sites (TSSs). We computed n-gram features are extracted and used in promoter prediction. We computed n-grams (n=2, 3, 4, 5) as features and created frequency features to extract informative and discriminative features for effective classification. Neural network classifiers with these n-grams as features are used to identify promoters in a human genome. Analysis of *n*-Gram based (AnG) is applied to the feature Extraction and a subset of principal components (PCs) are selected for classification. Our system uses three neural network classifiers to distinguish promoters versus exons, promoters versus introns, and promoters versus 3' un-translated region (3'UTR). We compared AnG-HPR with four well-known existing promoter prediction systems such as DragonGSF, Eponine and FirstEF, PCA-HPR. Validation shows that AnG-HPR achieves the best performance with three test sets for all the four predictive systems.

Keywords— promoter recognition; sequence feature; transcription start sites, Biological data sets, neural networks, Binary classification, cascaded classifiers.

1. INTRODUCTION

Promoter recognition is a real problem that computationally identifies the transcription start site (TSS) or the 5'end of the gene without time-consuming and expensive experimental methods that align ESTs, cDNAs or mRNAs against to the entire genome. In this article, we focus on humans because it is a representative species that has attracted much more attention in the past decade. In humans, the TSS is surrounded with a core-promoter region within around ±100 base pairs (bp). A proximal promoter region has several hundreds bp immediately upstream of the core promoter.

.

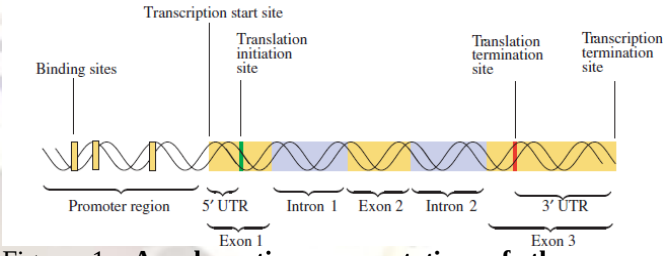

Figure 1: **A schematic representation of the locations of the promoter region, TFBSs, exons, introns and 3′utr regions.**

Available promoter prediction systems use two types of features for classification namely, context features like nmers, and signal features such as TATA-box, CCAAT-box, and CpG islands. Among the favorable promoter prediction programs, Eponine **[1]** builds a PWM to detect TATA-box and G+C enrichment regions as promoter-like regions; FirstEF **[2]** uses CpG-related and non-CpG related first exons as signal features; Promoter Inspector **[3]** uses IUPAC words with wildcards as context features. Good experiment results are achieved by integrating these two types of features. DPF **[4]** applies a separate module on G+C rich and G+C poor regions, and selects 256 pentamers to generate a PWM for prediction. Furthermore, DragonGSF **[5, 6]** adds the CpG-island feature to DPF.

Most of the promoter prediction programs try to predict the exact location of the promoter region of the known protein-coding genes, while some focus on finding the transcription start site (TSS). Some research has show that there is often no single TSS, but rather a whole transcription start region (TSR) containing multiple TSSs that are used with different frequencies. Generally two main approaches are used in promoter prediction [7].

1. First approach assigns scores to all single nucleotides to Identify TSS.
2. Second approach identifies a promoter region without providing scores for all nucleotides.

In this article analyzes the performance of 17 programs on two tasks: Genome wide identification of the start of genes and Genome wide identification of TSRs. In Existing methods Promoter prediction programs DragonGSF predicts the value of precision is between 37-48% and recall is 51-70%, DragonPF predicts the value of precision is between 61-65% and recall 62-64%, FristEF

predicts the value of precision is 79-81% and recall 35-40%, Eponine predicts the value of precision is ≈ 40 and recall ≈ 67%, NNPP2.2 predicts the value of precision is 69-93% and recall 2.0-4.5%, McPromoter2.0 predicts the value of precision is 26-57%and recall ≈ 4.5, proSom predicts the value of precision is 0.38% and recall 0.66% [8].
We select the 13 representative PPPs that can analyze large genome sequences and report strand-specific TSS predictions. ARTS, Eponine used SVM as the part of their design, EP3, Promoter scan, Wu-method used Position weight matrix (PWM).

CpGcluster used distance based algorithm. CpGProD used linear discriminating analysis (LDA). DragonGSF, DragonPF, McPromoter used neural networks.NNPP2.2 has used Time Delay neural network. Promoter Explorer has used AbaBoost algorithm. proSOM has used SOM. These programs have used as features various aspects of promoter and other regions of the DNA [6].
There are three classification methods for human promoter recognition system [9].
• Discriminative model that finds the optimal thresholds or classification boundaries in the signal, context and structure features space. Typical methods include artificial neural networks (ANNs), discriminate functions and support vector machines (SVMs).
• Generative model that describes the generative process of signal, context and structure observations. Position weight matrix (PWM), nearest neighborhood and hidden Markov models (HMMs) belong to generative models.
• Ensemble that combines multiple classifiers for multiple features in order to achieve a consensus and robust recognition results.

Promoter Inspector is a program that predicts eukaryotic poly II promoter regions with high specificity in mammalian genomic sequences. The program Promoter Inspector focuses on the genomic context of promoters rather than their exact location. Promoter Inspector is based (refer table 1) on three classifiers, which specialize in to differentiating between promoter region and a subset of non-promoter sequences (intron, exon and 3′utr).

In contrast to this, PromnFD and PromFind use only one classifier, i.e the features are extracted from one promoter set and one set of various non-promoter sequences. To compare the two approaches, two versions of Promoter Inspector are built. Version v1 was based on one st of mixed non-promoter sequences, while version v2 was built on the basis of exon, intron and 3′utr. Both versions of promoter Inspector were applied to exon, intron, 3′utr and promoter evaluation sets [3]. The identification of promoters and first exons has been one of the most difficult problems in gene-finding. The FirstEF [2] program identifies a promoter region and first exons in the human genome, which may be also be useful for the annotation of other mammalian genomes. FirstEF consists of different discriminate functions structured as a decision tree. The probabilistic models are designed to find potential first splice donor sites and CpGrelated and non-CpG-related promoter regions based on discriminant analysis. For every potential first splice-donor site and upstream promoter region, FirstEF decides whether the intermediate region could be a potential first exon based on a set of quadratic discriminant functions. Training and testing using different discriminant functions, the first exons and promoter regions from the first-exon database are used. Ramana et al. have also tried to identify the promoters as well as first exons of human species by using an algorithm called FirstEF which is based upon the usage of structural and compositional features [3]. They were able to predict 86% of the first exons. They have compared their method with Promoter Inspector and obtained a sensitivity of 70% compared to Promoter Inspector's 48%. Bajic et al. termed that the prediction is positive if the predicted transcription start site (TSS) falls within a maximum allowed distance from the reference transcription start site [5]. They have assessed performance of some of the prediction algorithms based on the performance measures such as sensitivity and positive predictive value. In their later paper they concluded that the promoter prediction combined with gene prediction yields a better recognition rate [6].

ExonType	Sensitivity	Specificity	correlation coefficient
CpG-related	0.92	0.97	0.94
NotCpG-related	0.74	0.60	0.65
all exons	0.86	0.83	0.83

Table1.Accuracy of FirstEF based on cross-validation

Li et al. is having proposed that locate eukaryotic promoter regions and predict transcription start sites (TSSs). Here the authors have computed codon (3-mer) and pentamer (5-mer) frequencies and created codon and pentamer frequency feature matrices to extract informative and discriminative features for effective classification. Principal component analysis (PCA) is applied to the feature matrices and a subset of principal components (PCs) are selected for classification [7].

Program	True Positive	False Positive	Sensitivity (%)	PPV (%)
DragonGSF	269	69	68.4	79.6
FirstEF	331	501	84.2	39.8
Eponine	199	79	50.6	71.6
PCA-HPR	301	65	76.6	82.2

Table 2. Performance comparison of four prediction systems for 22 chromosomes.

Promoter prediction systems use two types of features for classification namely, context features like n-mers, and signal features such as TATA box, CCAAT-box and CpG islands. Among the favorable promoter prediction programs, Eponine builds a PWM to detect TATA-box and G+C enrichment regions as promoter-like regions; FirstEF uses CpGrelated and non-CpG related first exons as signal features; Promoter Inspector uses IUPAC words with wild cards as context features. Good experiment results are achieved by integrating these two types' features. DPF applies a separate module on G+C rich and G+C poor regions, and selects 256 pentamers to generate a PWM for prediction. Furthermore, DragonGSF adds the CpG-island feature to DPF. Jia Zeng et al. in their paper selected three promoter systems DragonGSF, Eponine and FirstEF to compare the performance on test set 1. A promoter region is counted as a true positive (TP) if TSS is located within the region, or if a region boundary is within 200bp 5′ of such a TSS. Otherwise the predicted region is counted as a false positive (FP). The test results of Eponine and FirstEF. On test set 2, we adopt the same evaluation method as DragonGSF when one or more predictions fall in the region of [2000, +2000] relative to a TSS, a TP is counted. All predictions which fall on the annotated part of the gene in the region are counted as FP [10]. Sobha.et.al is termed that n-gram based promoter recognition methods were tried in promoter prediction and its application to whole genome promoter prediction in E.coli and Drosophila [11]. Here, we describe a method named AnG-HPR to predict the location of the TSSs with best performance. We extracting n-grams and using them to identify promoters in human genome. Patterns or features that characterize a promoter/non-promoter are needed to be extracted from the given set of promoter and non-promoter sequences. Here promoter recognition is addressed by looking at the global signal characterized by their frequency of occurrence of n-grams in the promoter region.

2. Introduction to N-Grams as features

Promoter recognition is tackled using various techniques such as support vector machines (SVM) [12], neural networks [13, 14], hidden Markov models [15], position weight matrix (PWM) [16], to expectation and maximization (EM) algorithm [17]. These techniques are based on basically motifs present in the promoter which are specific regions in the promoter or the global signal that is present in the promoter. To extract the local or global signals various feature extraction methods are used.

Condon usage patterns in coding regions and hexamer conservation (TATA box, CAAT box) in promoter regions is well known. Techniques that use these concepts are available in abundance in literature. Most of the local content-based methods are in fact based on conservation of the hexamers [13, 16]. In literature there are a few articles on protein sequence classification and gene identification using *n*-grams, but very few on promoter recognition. An *n*-gram is a selection of *n* contiguous characters from a given character stream [18]. Ohler *et al.* have used interpolated Markov chains on human and *Drosophila* promoter sequences as positive data set achieving a performance accuracy of 53% [19]. Ben-gal *et al.* have used a variable-order Bayesian network which looks at the statistical dependencies between adjacent base pairs to achieve a true positive recognition rate of 47.56% [20]. Leu *et al.* have developed a vertebrate promoter prediction system with cluster computing extracting *n*-grams for $n = 6$ to 20 [21]. They achieved an accuracy rate of 88%. Ji *et al.* have used SVM and *n*-grams ($n = 4, 5, 6, 7$) for target gene prediction of *Arabidopsis* [22]. Prediction system with cluster computing extracting *n*-grams for $n = 6$ to 20 [21]. They achieved an accuracy rate of 88%. Ji *et al.* have used SVM and *n*-grams ($n = 4, 5, 6, 7$) for target gene prediction of *Arabidopsis* [22]. There are position specific *n*-gram methods by Wang *et al.* and Li *et al.* [23, 24]. Wang *et al.* have proposed a position specific propensity analysis model (PSPA) which extracts the propensity of DNA elements at a particular position and their co-occurrence with respect to TSS in mammals. They have considered a set of top ranking *k*-mers ($k = 1$ to 5) at each position ±100 bp relative to TSS and the co-occurrence with other top ranking *k*-mers at other downstream positions. PSPA score for a sequence is computed as the product of scores for the 200 positions of ±100 bp relative to TSS. They found many position-specific promoter elements that are strongly linked to gene product function. Li *et al.* too have considered position-specific weight matrices of hexamers at some ten specific positions for the promoter data of *E. coli* [24].

Here, we extract *n*-grams to be used as features. An investigation of the lower order *n*-grams for promoter recognition was taken up in order to assess their applicability in whole genome promoter recognition. To this end we have extracted the *n*-grams and fed them to a multi-layer feed-forward neural network. Further, by using the best

features, two more neural networks are designed to annotate the promoters in a genome segment. In the following sections, we explain the extraction of features, the classification results using these features, and a way of finding the promoters in an unknown segment of the *Human* genome.

2.1 Feature Extraction

In this section, different data sets that are used in experimentation and the feature extraction method for various *n*-grams are described.

2.1.1 Data set

The training set in this experiment is divided into several subsets of promoters, introns, exons and 3′UTR sequences. Promoter sequences are extracted from two public databases. One is the Eukaryotic Promoter Database (EPD), release 86 **[8]**, which contains 1871 human promoter sequences. The other is the Database of Transcription Start Sites (DBTSS), version 5.2.0 **[9]**, which includes 30,964 human promoter sequences and 15,531 forward strand promoter sequences. We used forward strand promoter sequences are in our experiment. Human exon and intron sequences are extracted from the EID **[10]**, and the first exons are not included in the exons training set. Human 3'UTR sequences are from the UTR database **[11]**. From DBTSS we have extracted promoter sequences [-250, +50] bp around the experimental TSS. DBTSS contains 24 chromosomes; each chromosome has 1000 nucleotide sequences. From EID and 3′UTR we have extracted non-promoter sequences of length 300 bp [27].

2.1.2 Method

Patterns or features that characterize a promoter/non-promoter are needed to be extracted from the given set of promoter and non-promoter sequences. Here promoter recognition is addressed by looking at the global signal characterized by the frequency of occurrence of *n*-grams in the promoter region. We show in the section neural network architecture and classification performance that these features perform well for prokaryotic as well as eukaryotic promoter recognition. To extract the global signal for a promoter, the frequency of occurrence of *n*-grams is calculated on the DNA alphabet *{*A,T,G,C*}*. The set of *n*-grams for $n = 2$ is 16 possible pairs such as AA, AT, AG, AC, TA, etc. and the set of *n*-grams for $n = 3$ are 64 triples such as AAA, AAT, AAG, AAC, ATA etc. Similarly *n*-grams for $n = 4, 5, 6$ are calculated. Let f_i^n denote the frequency of occurrence of the *i*-th feature of *n*-gram for a particular *n* value and let $|L|$ denote the length of the sequence. The feature values V_i^n are normalized frequency counts given in Eq. (1).

$$V_i^n = \frac{f_i^n}{|L| - (n-1)}, \quad 1 \le i \le 4^n \text{ for } n=2, 3, 4, 5 \qquad (1)$$

Here, the denominator denotes the number of *n*-grams that are possible in a sequence of length $|L|$ and hence V_i^n denotes the proportional frequency of occurrence of *i*-th feature for a particular *n* value. Thus each promoter and non-promoter sequence of the data set is represented as a 16-dimensional feature Vector $(V_1^2, V_2^2, V_3^2, \ldots\ldots V_{16}^2)$ for $n=2$, as a 64-dimensional feature vector $(V_1^3, V_2^3, V_3^3, \ldots\ldots V_{64}^3)$ for $n=3$, as a 256 dimensional feature vector $(V_1^4, V_2^4, V_3^4, \ldots\ldots V_{256}^4)$ for $n=4$ and a 1024-dimensional feature vector $(V_1^5, V_2^5, V_3^5, \ldots\ldots V_{1024}^5)$ for $n=5$.

In a binary classification problem, the training set will be a mixture of both positive and negative data sets. Similarly the test set which consists of both positive and negative data is used to evaluate the performance of the classifier. A neural network classifier is trained using the *n*-grams of the training set as input feature vectors and then the test set is evaluated using the same network. Figures 1–4 depict the average separation between the positive and negative data for $n = 2$, $n = 3$, $n = 4$ and $n = 5$, respectively. It can be observed that the plots depict the separability of promoter and non-promoter data sets in different feature spaces.

CpG islands, which exist in 60% of mammalian promoters **[12]**, are regarded as one of the most important signal features in promoter recognition. Methods such as CpGProD **[13]**, DPF **[4]**, DragonGSF **[5, 6]**, FirstEF **[2]** and PromoterExplorer **[7]** embed this signal feature in their prediction system. In our method, two features are used to identify if a given sequence (>200bp) is CpG islands related: GC percentage (GCp) and Observed/expected CpG ratio (o/e). These are calculated with equations (1-5) (see supplementary material). If GCp > 0.5, and o/e > 0.6, then the sequence is considered CpG islands related, otherwise it is non-CpG islands related **[14]**. A DNA sequence contains four types of nucleotides: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). With different combinations, there are 43=64 codons and 45=1024 pentamers in promoter and nonpromoter datasets. Pentamers are widely selected as context features in many promoter prediction models **[4, 7]**, as they keep good balance between search efficiency and discriminability. Four resulting feature matrices are constructed from promoter, exon, intron and 3'UTR training datasets. Finally, three pairs of matrices: promoter versus exon, promoter versus intron and promoter versus 3'UTR are built from the four feature matrices for further processing.

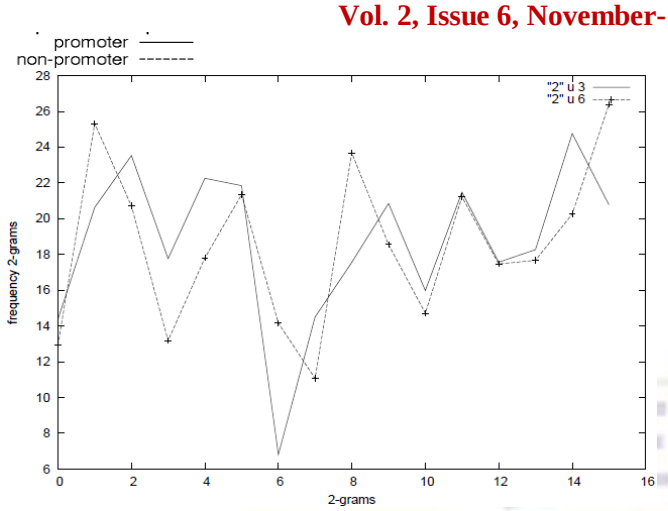

Fig.1. Average separation between promoters and non-promoters for *n* = 2 for *Homo sapiens*. Here, 1. . .16 represent AA, AT, AG, AC etc.

3. NEURAL NETWORK ARCHITECTURE AND CLASSIFICATION PERFORMANCE

A feed-forward neural network with three layers, namely, an input layer, one hidden and an output layer is used for promoter classification. The number of nodes in the input layer is 16, 64, 256, 1024 features for *n* = 2, 3, 4, 5 respectively. One more experimentation which uses Euclidean distance measure to reduce the number of features of *n* = 5 is done. Experimentation is done with different number of hidden nodes that give an optimal classification performance. The output layer has one node to give a binary decision as to whether the given input sequence is a promoter or non-promoter. 5-fold cross-validation [25] is used to investigate the effect of various *n*-grams on promoter classification by neural network. Average performance over these folds is being reported. These simulations are done using Stuttgart Neural Network Simulator [26].

S. No	Features	Precision	Specificity	Sensitivity	PPV
1	N=2 gram	68.47	84.05	67.36	83.51
2	N=3gram	70.86	81.923	63.94	84.89
3	**N=4gram**	**72.42**	**86.51**	**84.38**	**89.24**
4	N=5gram	76.56	69.44	80.85	81.54

Table 3. *Homo sapiens* classification results for different *n*-grams (average of 5-fold cross validation experiments)

3.1 Classification Performance

In a binary classification problem the training set will be a mixture of both positive and negative data sets. Similarly the test set also consisting of both positive and negative data is used to evaluate the performance of classifier. A neural network classifier is trained using n-grams of training set as input feature vectors and then the test set is evaluated using by same network. The below figures are depict the average separation between the positive and negative for n=2, 3, 4, 5 respectively. It can be observed that the plots depict the separability of promoter and non-promoter data sets in different feature spaces.

A feed forward neural network with three layers is used for promoter classification. The nodes in the input layer are 16, 64, 256, 1024 features for n=2, 3, 4, 5 respectively. The experimentation is done with different number of hidden nodes that give an optimal classification performance. The output layer has one node to give a binary decision as to whether the given input sequence is a promoter or non-promoter. 5-fold cross validation is used to investigate the effect of various n-grams on promoter classification by neural network. Average performance over these folds is being reported. These simulations are using Stuttgart Neural Network Simulator (SNNS). The classification results are evaluated using performance measures such as Precision, Specification, Sensitivity 5.3, 5.4 given these results.

Using these, we would like to find out the efficacy of these features in identifying promoter in human genome [2].

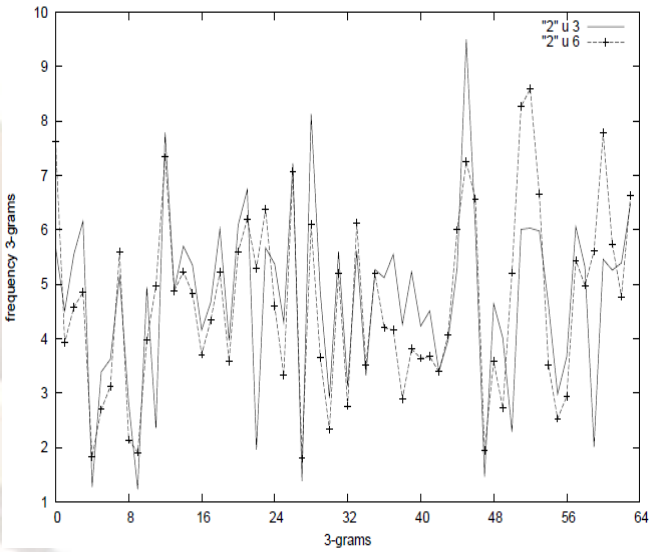

Fig. 2. Average separation between promoter and non promoter for n=3. Here 0...64 represent the AAA, AAT, AAG, AAC...etc

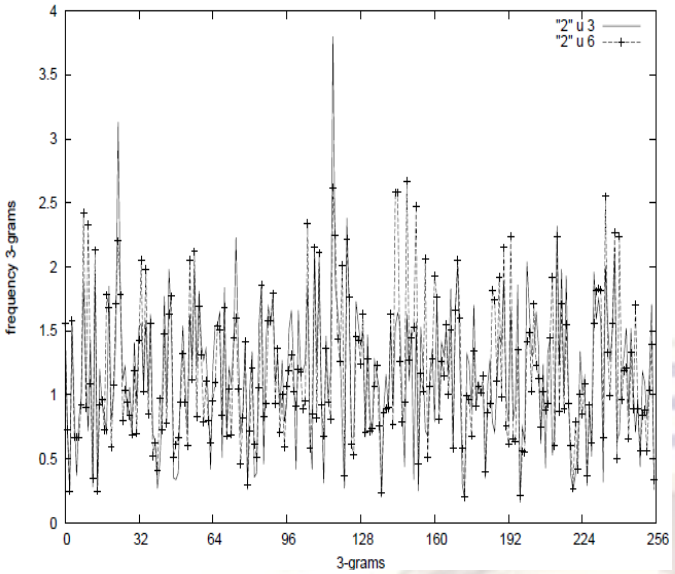

Figure.3: Average separation between promoter and non promoter for n=4. Here 0...256 represent the AAAA, AAAT, AAAG, AAAC...etc

n-gram	Precision	Specificity	Sensitivity	PPV
3-gram	82.07	82.86	78.06	83.75
4-gram	**82.51**	**84.95**	**78.64**	**85.51**

Table 4. *Homo sapiens* classification results for different *n*-grams for reduced data set (average of 5-fold cross-validation experiments)

Program	True Positive	False Positive	Sensitivity (%)	PPV (%)
DragonGSF	269	69	68.4	79.6
FirstEF	331	501	84.2	39.8
Eponine	199	79	50.6	71.6
PCA-HPR	301	65	76.6	82.2
AnG-HPR	431	61	78.64	85.51

Table 5. Performance comparison of five prediction systems for test set 2.

The classification results are evaluated using performance measures such as precision, specificity and sensitivity. Specificity is the proportion of the negative test sequences that are correctly classified and sensitivity is the proportion of the positive test sequences that are correctly classified. Precision is the proportion of the correctly classified sequences of the entire test data set.

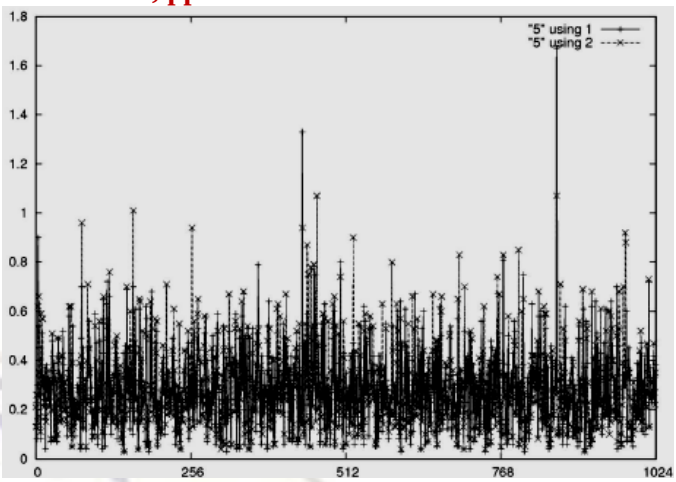

Fig.4. Average separation between promoter and non promoter for n = 5. Here 0...1024 represent the AAAAA, AAAAT, AAAAG, AAAAC...Etc

Positive predictive value is defined as the proportion of true positives to the sum of true positives and false positives. False positives are negatives which are classified as positives. The classification results for various n-grams for Homo *sapiens* presented in Tables 1 and 2 respectively.

4. DISCUSSION AND CONCLUSION

In proposed approach human promoters are classified using a neural network classifier. Since the optimal features are not known we started classification model with minimum number of features n=2 and incrementally increased to maximum number of features. The features are given as input to a single layer feed forward neural network. We selected four promoter systems, DragonGSF, Eponine and FirstEF, PCA-HPR to compare the performance on test set 1. A promoter region is counted as a true positive (TP) if TSS is located within the region, or if a region boundary is within 200bp 5' of such a TSS. Otherwise the predicted region is counted as a false positive (FP). The test results of Eponine and FirstEF are cited from the reference paper **[16]**. On test set 2, we adopt the same evaluation method as DragonGSF**[5]**: when one or more predictions fall in the region of[−2000, +2000] relative to a TSS, a TP is counted. All predictions which fall on the annotated part of the gene in the region [+2001, EndofTheGene] are counted as FP. Other predictions are not considered in counting TP and FP. Experimental results of DragonGSF, FirstEF and Eponine are obtained from **[5]**. We adopt the sensitivity (SN) and the positive predictive value (PPV) to evaluate the performance of these systems. The results and comparisons based on test 1 and test 2 are shown in Table 4 and Table 5.

In test 1, with the same number of true positives in comparison with existing methods, our method produces the smallest number of false positives. In test 2, although FirstEF, PCA-HPR

achieves a higher SN than AnG-HPR, the PPV is just half of AnG-HPR. PCA-HPR keeps a good balance between SN and PPV, while AnG-HPR produces better results. On test set 3, we compare AnG-HPR with DragonGSF because DragonGSF is the only online system which can accept relatively longer sequences among systems compared in the analysis. In order to get fair results for these sequences which are longer than 1,000,000bp (the limitation of a file in the DragonGSF web tool), we divided them into segments that are equal or less than 1,000,000bp each, before sending them to PCA-HPR and DragonGSF. Under the same evaluation criteria as the one in test set 2, AnG-HPR achieved a better result: the SN of AnG-HPR and DragonGSF are 76.2% and 46.8%, and PPV of the two systems are 83.4% and 63.8%, respectively. DragonGSF reports a good prediction performance on the whole human genome sequence, but it uses the TRANSFAC **[17]** database which includes binding site information only available for known promoters. Therefore, our system has the advantage in predicting unknown promoters. Different numbers of feature values are used to arrive at the best performance. Test results are measured using measures such as precision, specificity and sensitivity and PPV. Maximum accuracy achieved using SNNS is 89.2% in the case of DBTSS data set.

In this Paper, we focus on extracting the statistical features. There is evidence of a statistical preference in terms of codon usage patterns in protein coding regions. The majority of promoter prediction methods available now directly extract a limited number of context features from sequences. Here we are not doing any feature selection and using the entire set of n-grams. In this paper Classification of both promoter (DBTSS data set) and non-promoter is best for n=4. We obtained a precision of 72.42%, specificity of 86.5%, and sensitivity of 84.3% and positive predictive value of 89.2% for this set. The result shows that for DBTSS n=4 gram features give the better performance than other n-grams. The results here consolidate the results obtained for Drosophila Melanogaster in the work done by Sobha et al. They obtained best performance results for n=4. Does this then make 4-grams as a special parameter for the eukaryotes is needed to be further investigated?

A study of the n-gram (n=2, 3, 4, 5) as features for a binary neural network classifier is done. In human genome 4-gram features give them an optimal performance with neural network classifier. The results show that the classification result 4-gram is better in identification of the promoter than the other n-grams. Human promoter classification gives the better accuracy results of 89.2%.

ACKNOWLEDGEMENTS

We would like to thank Dr. T. Shoba Rani for her valuable comments in our discussions. We would like to thanks Dr.Bapi raju for his valuable comments in our discussion.

REFERENCES

[1] T. A. Down *et al.*, *Genome Research*, 12: 458 (2001) [PMID: 11875034].
[2] R. V. Davuluri *et al.*, *Nature Genetics*, 29: 412 (2001) [PMID: 11726928]
[3] M. Scherf *et al.*, *Journal of Molecular Biology*, 297: 599 (2000) [PMID: 10731414]
[4] V. B. Bajic *et al.*, *Journal of Molecular Graphics and Modelling*, 21: 323 (2003) [PMID: 12543131]
[5] V. B. Bajic *et al.*, *Genome Res.*, 13: 1923 (2003) [PMID: 12869582]
[6] V. B. Bajic *et al. Nature Biotechnology*, 22: 1467 (2004) [PMID: 15529174]
[7] Thomas Abeel, Yves van de peer, and Yvan Saeys. Towards a gold standard for promoter prediction evolution. *Bioinformatics*, 25(53):i313–i320, July 2009.
[8] Bajic,V.B. *et al*. (2006) Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.*, S3.1–S3.13.
[9] Jia Zeng, shanfeng Zhu, and Hong Yan. Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. *Briefings in Bioinformatics*, 10(05):498–508, 2009.
[10] Xiaomeng Li, Jia Zeng, and Hong Yan. A principle component analysis model for human promoter recognition. *Bio information*, 2(9):373–378, june 2008.
[11] Sobha Rani T and Raju S.Bapi. Analysis of n-gram based promoter recognition methods and application to whole genome promoter prediction. *In Silico Biology*, 9(1-2):S1–S16, March 2009.
[12] Gordon, L., Chervonenkis, A. Y., Gammerman, A. J., Shahmurradov, I. A. and Solovyev, V. V. (2003). Sequence alignment kernel for recognition of promoter regions. Bioinformatics 19, 1964-1971.
[13] Ma, Q., Wang, J. T. L., Shasha, D. and Wu, C. H. (2001). DNA sequence classification via an expectation maximization algorithm and neural networks: a case study. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, Special Issue on Knowledge Management. **31**, 468-475.

[14] Wang, J., Ungar, L. H., Tseng, H. and Hannenhalli, S. (2007). MetaProm: a neural network based meta-predictor for alternative human promoter prediction. BMC Genomics **8**, 374.

[15] Pedersen, A. G., Baldi, P., Chauvin, Y. and Brunak, S. (1999). The biology of eukaryotic promoter prediction - a review. Comput. Chem. **23**, 191-207.

[16] Huerta, A. M. and Collado-Vides, J. (2003). Sigma70 promoters in *Escherichia coli*: Specific transcription in dense regions of overlapping promoter-like signals. J. Mol. Biol.**333, 261-278**

[17] Cardon, L. R. and Stormo, G. D. (1992). Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. J. Mol. Biol. **223**, 159-170.

[18] Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal. **27**, 379-423.

[19] Ohler, U., Liao, G. C., Niemann, H. and Rubin, G. R. (2002). Computational analysis of core promoters in the *Drosophila* genome. Genome Biol. **3**, research0087.

[20] Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S. and Grosse, I. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics **21**, 2657-2666.

[21] Leu, F., Lo, N. and Yang, L. (2005). Predicting Vertebrate Promoters with Homogeneous Cluster Computing. Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems, SITIS. 143-148.

[22] Ji, H., Xinbin, D. and Xuechun, Z. (2006). A systematic computational approach for transcription factor target gene prediction. IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology CIBCB '06, 1-7.

[23] Wang, J. and Hannenhalli, S. (2006). A mammalian promoter model links cis elements to genetic networks. Biochem. Biophys. Res. Commun. **347**, 166-177.

[24] Li, Q. Z. and Lin, H. (2006). The recognition and prediction of σ70 promoters in *Escherichia coli* K-12. J. Theor. Biol. **242**, 135-141.

[25] Alpaydin, E. (2004). Introduction to Machine Learning, MIT Press.

[26] Stuttgart Neural Network Simulator http://www-ra.informatik.uni-tuebingen.de/SNNS/

[27] Benchmark datasets. www.fruitifly.org/data/seq-tool/datasets.com.