Publications


Promoter Prediction using Horspool’s Algorithm

T. Uma Devi¹, P.V. Nageswara Rao², Prof. Allam Appa Rao³
¹Dept. of Computer Science, ²Dept. of Computer Science & Engineering, GITAM University, Visakhapatnam.
³Dept. of Computer Science & Systems Engineering, Andhra University, Visakhapatnam.

Abstract - A promoter, the region in DNA encompassing a gene’s transcription start site (TSS), largely controls the biological activation of the gene. The computational promoter prediction has been based on specific genomic context that promoter has in large genomic sequence. We aimed to detect consensus sequences of promoter i.e. TATA box, INR, CAAT box and GC box. The consensus sequence is an ideal sequence for the interaction with its regulatory protein. We propose an algorithm for consensus sequence matching called Horspool’s Algorithm and extensions to this algorithm is to identify the position where these consensus sequences occur and give a score based solution on the identification of all these 3 or 2 or 1 consensus sequences. The promoter is recognized based on this score.

Keywords: Promoter Prediction, Consensus Sequence, Horspool’s Algorithm

1 Introduction

One of the challenges in the field of Computational Biology and especially in the area of computational DNA sequence analysis is the automatic detection of promoter sites. As a result of the ongoing Human Genome Project, DNA or protein data are accumulated at a speed growing at an exponential rate. One of the main goals of the Human Genome Project is the characterization, annotation-recognition and categorization of genes from human genome to serve as a periodic table for biomedical research (Lander, 1996). The process of transcription begins with the RNA polymerase (RNAP) binding to DNA in the promoter region, which is in the immediate vicinity of the transcription start site (TSS). Exactly, how RNAP locates this specific binding site in the large excess of non-promoter DNA remains a field of intense investigation. A typical promoter sequence is thought to comprise some sequence motifs positioned at specific sites relative to the TSS.

2 Transcription

Initiation of transcription is the first step in gene expression, and constitutes an important point of control in prokaryotes as well as in eukaryotes (Reznikoff et al. 1985). Transcription initiates when RNA polymerase recognizes and binds to certain DNA-sequences termed promoters. Subsequent to binding, a short stretch of DNA double helix is disrupted, and the polymerase starts to synthesize RNA by the process of complementary base pairing. The sequence of the promoter determines the position of the transcriptional start point and is further more important for the frequency with which the gene is transcribed (the strength of the promoter). Our goal is to identify promoters in the unlabeled DNA sequences. This task is particularly difficult in the case of eukaryotic organisms in which regulatory regions represent a small percentage, overwhelmed by presumably non-functional DNA. So prediction and characterization of regulatory regions is still a challenging problem. Here we focus on detecting promoters, which are in the class of regulatory regions. Extracting the complete functional information encoded in a genome is a central challenge in biological research. Prediction of non-protein coding functional regions, such as regulatory elements, is especially difficult because of usually short (6–15 bp for S. cerevisiae and many other eukaryotic genomes), often degenerate, and can reside on either strand of DNA at variable distances from the genes they control. Since functional sequences tend to be conserved through evolution, they can appear as ‘phylogenetic footprints’ in alignments of genome sequences of different species (Hardison et al., 1997).

3 Promoter Recognition

Prediction of a promoter is one of the many active areas of bioinformatics. A promoter should contain an element which is identical to or very close to the consensus sequence. The outcome of a promoter detection algorithm is directly or indirectly influenced by the success of identifying the location of a TATA box, in a promoter sequence. In prokaryotes, the sequence of a promoter is recognized by the Sigma (σ) factor of the RNA polymerase. In eukaryotes, it is recognized by specific transcription factors.

3.1 E-Coli Promoters

E.coli promoters has led to the identification of three major conserved features: the “-10 box”, the “-35 box”, and a pyrimidine (C or T) followed by a purine (A or G) at the initiation site (Rosenberg & Court 1979; Hawley & McClure 1983). The -10 and -35 boxes are conserved hexa nucleotide elements that are named according to the approximate position of their central nucleotides relative to the transcriptional start point. The well known consensus sequences are TTGACA for the -35 box, and TATAAT for the -10 box.

<table>
<thead>
<tr>
<th>σ Factor</th>
<th>Promoter Consensus Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>TTGACA</td>
</tr>
<tr>
<td>0.34</td>
<td>TCTGCCCCTTGAA</td>
</tr>
<tr>
<td>0.28</td>
<td>CTGAAA</td>
</tr>
<tr>
<td>0.21</td>
<td>CTGGNA</td>
</tr>
</tbody>
</table>

* -10 region is also called Pribnow box, after its discover.
+ N = any
3.2 Eukaryotic Promoters

![Diagram of Eukaryotic Promoter](image)

In Eukaryotes a promoter, usually appears near the beginning of the gene whose transcription it regulates. Promoters generally indicate and contain the starting point of transcription, the TSS, and they regulate the rate of initiation of transcription. The transcription is achieved through special proteins, enzymes called RNA polymerase II, that bind to certain parts of the DNA (promoter regions) and start 'reading' and storing in a mRNA sequence each gene code. RNA polymerase II is responsible for the synthesis of all mRNA.

In eukaryotes, there is a significant difference between the transcription of protein genes and RNA genes. The most common promoter element in eukaryotic protein genes is the TATA box, located at -35 to -20. Its consensus sequence, TATAAAA, is quite similar to the -10 region of the E.coli recognition site. Another promoter element is called the initiator (Inr). It has the consensus sequence PyPyAN(T/A)PyPy, where Py denotes pyrimidine (C or T), N = any, and (T/A) means T or A. The base A at the third position is located at +1 (the transcriptional start site). TATA box and initiator are the core promoter elements. There are other elements often located within 200 bp of the transcriptional start site, such as CAAT box and GC box which may be referred to as promoter-proximal elements.

**Eukaryotic Promoter elements**

<table>
<thead>
<tr>
<th>Promoter</th>
<th>Position</th>
<th>Transcription Factor</th>
<th>Consensus Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATA box</td>
<td>-35–20</td>
<td>TBP</td>
<td>TATAAAA</td>
</tr>
<tr>
<td>CAAT box</td>
<td>-200–70</td>
<td>CBF, NFI, C/EBP</td>
<td>CCAAT</td>
</tr>
<tr>
<td>GC box</td>
<td>-50–70</td>
<td>SP1</td>
<td>GGCGGG</td>
</tr>
</tbody>
</table>

* Most, but not all, CAAT and GC boxes are located between -200 and -70.
+ CBF = CAAT binding protein; C/EBP = CAAT/EBP binding protein.
++ N = any (A, T, G, or C); Py = Pyrimidine (C or T).

The protein which interacts with the initiator and TATA box is known as the TATA-binding protein (TBP), because the TATA box was discovered earlier than the initiator. TBP recognizes not only the core promoter of protein genes, but also RNA promoters. The TATA box is important in determining the precise starting point for transcription and two other promoter elements are the CAAT box (GGCCAATCT) and GC box (GGCGGG) located in the -50 to -100 region of structural genes in eukaryotes. In addition it has some transcription factor binding sites.

4. Horspool’s Technique

The problem of string matching requires finding an occurrence of a given string of m characters in a longer string of n characters. Starting with the last character of the pattern (say TATA box or Inr or GC box) and moving right to left, we compare the corresponding pairs of characters in the pattern and the sequence. If all the characters match successfully we say the position where the string occurred and specify the score. If we encounter a mismatch we shift the pattern right. We would like to make as large shift as possible without risking the possibility of missing a matching substring in the text. This algorithm determines the size of such a shift by looking at the character c of the text that was aligned against the last character of the pattern.

4.1 Preparation of Shift Table

Horspool’s algorithm determines the size of such a shift by looking at the character 'c' of the text that was aligned against the last character of the pattern. The shift table is prepared for the pattern based on 4 cases:

**Case 1:** If there is no character say 'c' in the pattern we will shift the pattern by its entire length.

**Case 2:** If there is occurrences of character say 'c' in the pattern but it is not the last one there then we shift the pattern such that the rightmost occurrence of 'c' in the pattern will align with the 'c' in the text.

**Case 3:** After some number of matching characters say 't' a mismatch found for the sequence character 'c' and there are no 'c's among its first 'm-t' characters in the pattern then we will shift the pattern by its entire length.

**Case 4:** After some number of matching characters say 't' a mismatch found for the sequence character 'c' and if 'c' happens to found among its first 'm-t' characters in the pattern then we shift the pattern such that the rightmost occurrence of 'c' in the pattern will align with the 'c' in the text.

4.2 Algorithm

1. For the characters used in both pattern and sequence we constructed the shift table. Shift Tables are prepared separately for each core and proximal promoter elements.

2. Align the pattern against the beginning of the sequence.

3. Starting with the last character in the pattern, compare the corresponding characters in the sequence until either all m characters are matched or a mismatching pair is encountered. If a mismatch found we retrieve shift value of the sequence character from the shift table and shift the pattern to the right along the sequence. We repeat this step until a matching...
string is found or pattern reached beyond the last character of the sequence.

We apply all the above 3 steps for each promoter element and maintain the score. If score exceeds the value $\geq 0.7$ we say it is the promoter.

From Promoter Database of E-coli (PromEC) and Eukaryotic Promoter Databases(EPD) especially of Drosophila Melanogaster(1926), Mus musculus (mouse) (196), Homo sapiens (man) (1871) retrieved the sequences and apply the algorithm for finding consensus sequence. The scores are : if the sequence consists of TATAAA sequence and INR is found within -26 to +32 range it is 0.8 if TATA box alone found with different nucleotide at +1 position the score is taken as 0.7. If it finds CAAT and GC Box then the score raised to 1.0. If one of them is found the score is 0.9. If the Score is $>0.9$ there exists promoter in the given sequence and if score $\geq 0.7$ there may be chance of promoter in the given sequence but if score $<0.7$ no promoter exists in the given sequence.

5. Conclusion

Main aim of this paper is to have faster identification of consensus sequence in a given portion of DNA sequence. The performance of searching can be improved through Horspool's algorithm. The problem with these consensus sequence in eukaryotic promoters is that they vary a lot; they may appear in different combinations, their relative locations with respect to the TSS are different for different promoters, and not all of these specific sub regions need to exist in a particular promoter. The high complexity of eukaryotic organisms led to specialization of the genes, so that promoters in eukaryotes are adjusted to their different conditions of expressions; for example in different cell types or tissues. Thus, the variability of internal eukaryotic promoter structure can be huge and the characteristics of the eukaryotic promoter are rather individual for the promoter than common for a larger promoter group. For this reason it is not easy to precisely define a promoter in eukaryotic organisms. And this is also one of the reasons why at this moment there is no adequate computer tool to accurately detect different types of promoters in a large-scale search through DNA databases. This algorithm can find promoter only if it has the consensus sequence and especially if it is having TATA box :[15] could not identify the TATA less promoters.

6. References


AN IN SILICO GENOME WIDE IDENTIFICATION, CHARACTERIZATION AND MODELING OF HUMAN PAPILLOMA VIRUS STRAIN 92

Dr. D.S.V.G.K.KALADHAR *, T. UMA DEVI, P.V. NAGESWARA RAO
GITAM University, Visakhapatnam-530045, India.

Abstract:
Most of the biologists focus to explore innovations of their research in faster rate using developments in Information technology. The gene identification, characterization and modeling of the proteins in HPV 92 is done using bioinformatics tools. A complete genome of HPV-92 with NCBI's accession number NC_004500 was submitted to FGENES V0, a viral gene prediction server, predicts six genes. These six genes are characterized as E6 oncoprotein, E7 oncoprotein, E1 Replication protein, E2 Regulatory protein, L1 major capsid protein and L2 minor capsid protein. Isoelectric points and Molecular weights of all the six proteins vary largely and the modeled structures are shown. The research can provide characterization and modeling of genome which can further implemented in drug designing methods using bioinformatics tools.

Keywords: Human papilloma virus-92, gene prediction, modeling.

1. Introduction

Information Technology and Biological sciences today is being transformed due to enormous growth of data from laboratories world wide. The challenge is to transform information into knowledge that will lead to a better understanding of the system processes underlying both health and disease.

The human papillomavirus (HPV) family is ubiquitous in the human inhabitants and more than 100 virus types have been identified. HPV are double stranded DNA viruses and genome size measures approximately 8kb in length. The viruses are all small and non-enveloped, the protein coat surrounding the genome is an icosahedron, or a polygon with 20 faces, that is 55- nm in diameter.

Throughout the past 20 years, the perception of cervical carcinoma has shifted from that of a mysterious, sexually transmitted disease to one intimately related to HPVs. This evolution in understanding has been driven by a potent relationship between virus and disease, and by a wealth of molecular data supporting mechanisms of papilloma virus mediated tumorigenesis. HPVS associated with the development of "common warts" can be one, transmitted environmentally or by casual skip to skin contact. HPV infection is an important factor in the development of nearly all cases of cervical cancer.

Approximately 20 million people are currently infected with HPV, typically invades the skin without any apparent symptoms. An infection by HPV is detected when the virus stimulates the growth of the skin into small outgrowth or warts. Severe HPV types have also been associated with an increased risk for other cancers, including other genital and lung cancers.

Information processing and information slow occur in the course of an organism's development and throughout its lifespan. Organism's do not exist in segregation, but interact with each other constantly with in a complex ecosystem. The relationships between organisms, such as those between prey and predator, host and parasite are complex and multidimensional. Bioinformatics can be used in all stages of genetics research and can improve
study design assist in candidate gene identification, to aid data interpretation and management, to shed cut to the molecular pathology of disease-causing mutations.

2. Methodology

A complete genome of HPV type 92 was selected from NCBI (National center for Biotechnology Information) with accession number NC_004500 for the present study. The total length of the double stranded circular DNA viral molecule was 7461 bp. The sequence was submitted to FGENES V0, a fastest and most accurate ab initio gene prediction program in Viral Genomes to predict total number of genes located and was transformed to the protein sequences. The protein sequence is aligned based on multiple sequence protein alignment program, BLAST P for characterization of different proteins at molecular level. The aligned sequence is compared with other HPV streams for evolution of agents with HPV-92. Molecular weight and Isoelectric points of these molecules were predicted using tools present at ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (http://expasy.org/tools/pi_tool.html). Modeling of protein molecule was done using Swissmodel server.

3. Results and Discussion

A complete gene of HPV-92 has been retrieved from NCBI site with accession number NC_004500. The nucleic acid of this organism was submitted to FGENES V0 predicts 6 genes in the complete genome of HPV 92. All these 6 translated genes is submitted to BLASTP for characterization and the predicted proteins has been shown in TABLE 1. These six genes are characterized as E6 oncoprotein, E7 oncoprotein, E1 Replication protein, E2 Regulatory protein, L1 major capsid protein and L2 minor capsid protein.

Isoelectric points and Molecular weights (Table 1) of all the six proteins vary with protein sequences. The replicative, regulatory and capsid proteins are having higher molecular weights compared with oncoproteins E6 and E7 of HPV-92. Except Regulatory protein, all the proteins acquired acidic pH. The molecular weight of proteins range from 10.11 KD to 70.33 KD. Isoelectric point ranges from 4.35 to 10.48.

Most biologists focus their research in Bioinformatics for using computers to store, compare, retrieve, analyze or predict the composition and the structure of biomolecules. “Biomolecules” include genetic materials such as nucleic acids and the products of genes and proteins (central dogma of life). These are the consensus of classical Bioinformatics, dealing primarily with Biological Databases, sequence analysis and molecular modeling.

Modeling of all these protein sequences has been designed using swissmodel and is presented in Table 2.
4. Tables

<table>
<thead>
<tr>
<th>S.NO</th>
<th>GENE</th>
<th>LOCATION</th>
<th>No of Residues</th>
<th>CHARACTERIZATION</th>
<th>Isoelectric point</th>
<th>Molecular weight in Daltons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Genel</td>
<td>106 - 522</td>
<td>138</td>
<td>E6 protein oncoprotein</td>
<td>6.29</td>
<td>15808.44</td>
</tr>
<tr>
<td>2</td>
<td>Genel2</td>
<td>519 - 794</td>
<td>91</td>
<td>E7 oncoprotein</td>
<td>4.35</td>
<td>10115.63</td>
</tr>
<tr>
<td>3</td>
<td>Genel3</td>
<td>784 - 2622</td>
<td>612</td>
<td>E7 oncoprotein</td>
<td>4.35</td>
<td>10115.63</td>
</tr>
<tr>
<td>4</td>
<td>Genel4</td>
<td>2564 - 3997</td>
<td>477</td>
<td>E2 Regulatory protein</td>
<td>10.48</td>
<td>54135.57</td>
</tr>
<tr>
<td>5</td>
<td>Genel5</td>
<td>4056 - 5627</td>
<td>523</td>
<td>E2 minor capsid protein</td>
<td>4.83</td>
<td>57557.16</td>
</tr>
<tr>
<td>6</td>
<td>Genel6</td>
<td>5638 - 7176</td>
<td>512</td>
<td>L1 major capsid protein</td>
<td>6.23</td>
<td>57740.11</td>
</tr>
</tbody>
</table>

Table 1. Characterization, pi and Mw of predicted proteins in HPV-92.

<table>
<thead>
<tr>
<th>S.NO</th>
<th>Protein</th>
<th>Sequence</th>
<th>Swissmodel (template)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E6 protein oncoprotein</td>
<td>MAKPFSVQELRQLHDPLEDILHHCNCEAfPLTFEELIGQDAKLNNLWKNKENYACACGACLQVAETKYPHEYSVQDKDEIRDSGLLCClTVhXFeKCLRLHDLYLEKLVACASGFPRVRGAWKAVCRFTCEI</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>E7 oncoprotein</td>
<td>MKGQXAITPDVLDDLQYQDILHCDCEDLSHINQEEEFAQFRIDYK1VSSCGGCQIKLRIFASCTQGRTILQDILIEEALLCIPDiCNGR</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E1 Replication protein</td>
<td>MAGKIDPGKEGCSTWCLHABDCSVDGDFFEKLFDKXTDSEDDDLDDGGDLDGAELNGQFEELCILQEREESDLQMQLEKLKYFPKAV1Q1LSPLQESITSPQSKRFQRLFEQDOSQELSISHAEDSV5EAVEPFG5KDDVPEYSAATBKGSNKHE HyKQCLOC5NSNARATLLSKFKAAP 5PSWFEFELRFLKVHDNCTDCADWQAANDQHEBCHKLLSELITNYLHYLHSPLLNLVLYLHPLCYLFKQPKSKPEKAVKR</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>E2 Regulatory protein</td>
<td>MEALSOFNALSQKLKMITYRASSELKLQDIEHWNKLRQEQVLFH1VARKONGACQPGVPAKLQVSAEAKAEMYSVQKLAMERDGKNEKWTJLNETSVEITRFTPENCFKKGKPK7IE1HYDINGFNFTMLTYT1WTHYFEDDEGNGWQKTEGHLHYDHYAFMDQLNGKYQYIRFAQDAA 1RFSRESETGEWYKFKNEFLLAPVYTSSTNESSEDRPAPATYDGSLSQSTGOSGQFVPTQKQINSQERGVYKXDDATTAQSG0R5PRSRS</td>
<td>No valid models produced</td>
</tr>
<tr>
<td>5</td>
<td>L2 minor capsid protein</td>
<td>MALKRTKRSSTYINRTY1CKAATCGTTPYKVEQFTYDQVADQLK 1YQ5STVFGPFGLGDTGEGTOSTYQYTGQFGLGETYQVFTRP 1ALVFEPFAFGSFIIPSDIPFSSAIHPPPESTSPGPDPLLEGEITA1EV1HPADFTTVDPYTVTGRSNSAIVLBEAVDPFTSPTNRVSRTQYNNP 1AFOXIGSETTFSAGETLSLDQYQYDGDQOQIGNyPPFSSPYVEILNE 1HP50YFSEEEFPPTPPGTS1TPVQAPQMAALAYNEF8FTQVQV1VENDMFPFTPFRSRLVRQFNDPFVPYEBTVYQVFRDLLEITEEPDFRQF 1LDVQKLRQFTYATFCAYRIKSLRKLKATRTS1TSQITAIG1QY5HFR1DIESDTPFSIELQVLGIEHGDATTVQIGQYFPSTVNI1DEELPNEEN 1VELESDLLRKLEINFLGQSAQPQFNGRSN5STTVITYLPRFTPEVSFLSY1TV1DLQH1SIEPRESAYPEVFTEPNTD1VH1TIEDFSGYDYLHFLS1K1WK1GKEAY11</td>
<td>No valid models produced</td>
</tr>
<tr>
<td>6</td>
<td>L1 major capsid protein</td>
<td>MSYWFLAVNGKVPVLLQPFVQZSTDFQYQRTFN1YHANSDRLLTVGHPYFETDQSTPSASPNIPLQH11DDCISQECPTQFPEMYLVPC1QP 1LVPC1QPFLGPEKNTVIDTGGDMQV1GQGNNKFLASVTZDVS</td>
<td>No valid models produced</td>
</tr>
</tbody>
</table>

Table 2. Modeling of protein sequences using Swissmodel.
Acknowledgments

The authors would like to thank GITAM University for providing computational facility and access to e-journals to carry out this research.

References


INSILICO PROMOTER PREDICTION USING GREY RELATIONAL ANALYSIS

UMA DEVI TATAVARTHI1, VENKATA NAGESWARA RAO PADMANABHUNI1, APPA RAO ALLAM2, RAMACHANDRA SRIDHAR GUMPENY3

1GITAM University, Visakhapatnam,
2JNTUK, Kakinada, 3Endocrine and Diabetes Centre, Visakhapatnam
E-mail of corresponding author: uma@gitam.edu1

ABSTRACT

In machine learning, multiclass or multi-label classification is the special case within statistical classification of assigning one of several class labels to an input object. The multiclass problem is more complex than binary classification and less researched problem. In biology promoter is the DNA region where the transcription initiation takes place. Reliable recognition of promoter region is essential for understanding biological mechanism of the gene. This study proposes a new approach for predicting the promoter from the DNA sequence based on the modeling of Grey Relational Analysis (GRA). In order to construct a promoter prediction system, GRA approach is developed and applied to the real data set with 2111 samples of promoters and non-promoters of 4 species. The results of the current model are compared to those of traditional ones, logistic regression and back-propagation neural network. The results illustrate that the prediction of the proposed GRA model demonstrates better prediction accuracy than the conventional ones. The current results show that the proposed GRA provides a novel approach in predicting the promoter from a genome.

Keywords: Promoter Prediction, Grey Relational Analysis, Grey Systems, Classification

1. INTRODUCTION

Promoters are modular DNA structures containing complex regulatory elements required for gene transcription initiation (Lin and Li, 2010). In genetics, a promoter is a region of DNA that facilitates the transcription of a particular gene. Promoters are typically located near the genes they regulate, on the same strand and upstream (towards the 5' region of the sense strand) (Promoter (Biology), 2007). Hence, the identification of promoters using machine learning approach (insilico) is very important for improving genome annotation and understanding transcriptional regulation. In recent years, many methods have been proposed for the prediction of eukaryotic and prokaryotic promoters. However, the performances of these methods are still far from being satisfactory.

Recent availability of several genome sequences has allowed whole genome analyses to unravel their functional properties. One of the challenges of the genomics is to understand how genomes are transcribed. Specifically, the genes are small sequences spread out along the genomes which, after being transcribed in mRNA, are translated in proteins turning out to be functional units of the cells. Although several sequences located within the genes or in their closed vicinity control their specificity of expression there are typical regions, the promoters, that define the Transcription Start Site (TSS) of the genes.

In this paper, a new approach is proposed for predicting the promoter using machine learning algorithm: Grey Relational Analysis. By applying the proposed method to the promoter and non-promoter sequences of Homo Sapiens, Drosophila Melanogaster, Escherichia Coli (commonly abbreviated as E. Coli) and Saccharomyces Cerevisiae, the sensitivities and specificities obtained are: 94.9% and 96.54% for Homo Sapiens, 96.4% and 100% for Drosophila Melanogaster, 93.2% and 100% for Escherichia Coli and 98.3% and 94.14% for Saccharomyces Cerevisiae. The high accuracies indicate that this method can be used as an efficient method for the identification of eukaryotic and prokaryotic promoters. This approach can also be extended to
predict other species promoters at genome level also.

2. RELATED WORKS:

To accurately predict promoter regions, finding discriminative and informative features is the first and key step. As far as feature choice is concerned, there are two distinct types of features used in the area of promoter prediction: signal and context structure features. The most important signal features include CpG islands, transcription factor binding sites (TFBSs) such as TATA-box and CAAT-box, and initiator (Inr). PWM (Bucher, 1990) derives four weight matrices of TATA-box, cap signal, CCAAT-box and GC-box respectively. PromoterScan (Prestridge, 1995) uses a weight matrix to score TATA-box. A weight matrix is a simple generative model for a short, ungapped sequence motif (Down and Hubbard, 2002). PWM is used extensively in signal feature extraction processing, as it can create a profile that represents the common feature across the training sequence. This profile can be used to scan new sequences and make a decision as to whether these sequences are related to the training group (Raychaudhuri, 2006).

Hidden Markov Model (HMM) (Krogh and Brown, 1994) is a more sophisticated method for feature extraction from sequences compared to PWM. HMM can represent spacer-included motifs (Murakami et al., 2000) of a sequence family. Generalized Hidden Markov Model (GHMM) (Stormo and Haussler, 1994) is used for generating multi-symbol strings in gene finding systems (Kulp et al., 1996). The Pol II promoter prediction program (Murakami, et al., 2000) is built based on PromFD (Chen et al., 1997) and utilizes HMM to acquire additional motifs.

McPromoter is developed based on GenScan (Burge and Karlin, 1997), and uses stochastic segment models (SSMs) (Ostendorf et al., 1995) which is a generalization of HMM to represent six segments of the promoter sequence from -250 to +50bp: upstream 1 and 2, TATA box, spacer, initiator and downstream (Uwe Ohler, 2006).

3. MATERIAL AND METHOD:

In order to accomplish the task of promoter prediction, positive datasets and negative datasets were taken from different curated databases available on the World Wide Web. These databases contain non-redundant collection of promoters, for which the transcription start site has been determined experimentally. Table 1 describes the database name, species and number of sequences in that class, that are used in this study.

<table>
<thead>
<tr>
<th>Class</th>
<th>Database</th>
<th>Taxonomic group/organism</th>
<th>No. of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EPD</td>
<td>Homo Sapiens</td>
<td>608</td>
</tr>
<tr>
<td>2</td>
<td>DCPD</td>
<td>Drosophila Melanogaster</td>
<td>192</td>
</tr>
<tr>
<td>3</td>
<td>PromEC</td>
<td>E.Coli</td>
<td>471</td>
</tr>
<tr>
<td>4</td>
<td>SCPD</td>
<td>Saccharomyces Cerevisiae</td>
<td>232</td>
</tr>
<tr>
<td>5</td>
<td>Essential Genes</td>
<td>Homo Sapiens</td>
<td>118</td>
</tr>
<tr>
<td>6</td>
<td>Essential Genes</td>
<td>Drosophila Melanogaster</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>Essential Genes</td>
<td>E.Coli</td>
<td>300</td>
</tr>
<tr>
<td>8</td>
<td>Essential Genes</td>
<td>Saccharomyces Cerevisiae</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>2111</td>
</tr>
</tbody>
</table>

Table 1: Class Label, Database, Taxonomic Group/organism, Number of Sequences considered

4. Feature Extraction from DNA sequence:

Good input representations make it easier for the classifier to recognize underlying regularities. Therefore, good input representations are crucial to the success of classifier learning. The sequence or primary structure of a nucleic acid is the exact specification of its atomic composition and the chemical bonds connecting those atoms. From a one dimensional point of view, a DNA sequence contains characters from the 4-letter nucleic acid alphabet A,C,G or T. Let a given DNA sequence either promoter or non-promoter be: $S = s_1s_2s_3......s_{2L_1}s_2$, where $s_1 \epsilon \{A,C,G,T\}$, $1 \leq i \leq L$. The feature values of the sequence are calculated on the composition of the tri-mers. i.e., $L = \{AAA, AAC, AAG, AAT, ACA, ACC, ..., TTT\}$. The 64 compositional frequencies are calculated as:

$$f_i = \frac{S_i}{|S|} - 2$$

$f_i$ denote the frequency of occurrence of the $i^{th}$ feature and $|S|$ denote the length of the sequence. The feature values $f_i$ are normalized frequency counts (Nageswara Rao et al., 2008). Feature extraction is transforming the data in the high-dimensional space to a space of fewer dimensions. By applying Principal Component Analysis on the 2111x64 matrix it is reduced to 2111x41. The
obtained matrix is considered as input for the GRA Classifier.

5. Grey Relational Procedure:

Grey Relational Analysis (GRA) has been one of the most practical analytical tools (Deng, 1988; Liu and Lin, 2005; Nagai and Yamaguchi, 2004; Wen, 2004; Wen et al., 2006; Yamaguchi et al., 2006b). Several GRA models are developed and well summarized in (Liu and Lin, 2005; Wen et al., 2006; Yamaguchi et al., 2007; You et al., 2006). The GRA models provide appropriate tools for examining a rank of order of multiple objects with resemblance from an objective. In the recent years, GRA models have been applied to a lot of applications, such as decision making in computer science (Akabane et al., 2005; Yamaguchi et al., 2006a), system modeling, social science, geometry, chemistry, management (Lin et al., 2009; Nagai et al., 2005; Rui and Wunshch, 2005; Yamaguchi et al., 2005; Yamaguchi et al. 2006a), economics, marketing research (Yamaguchi et al., 2004).

Owing to the usefulness and robustness of GRA, a new approach for predicting the promoter is proposed, based on GRA. This study applies GRA model to predict promoter from the data set of 2111 samples from 4 species. The results of the current model are compared to those of traditional and hybrid models, which include conventional logistic regression, logarithm logistic regression and ANN approaches. The result shows that in predicting the promoter, GRA model gave better performance and demonstrates stronger prediction power than conventional logistic regression and ANN approaches. A system that has no information is defined as a black system, while a system that is full of information is called white. Thus, when the information of a system is either incomplete or undetermined, it is defined as grey system. The grey number in grey system represents a number with incomplete information. The grey element represents an element with incomplete information. The grey relation is the relation with incomplete information. This section describes the basic definitions of grey relational analysis, GRA. The inner product and metric of two vectors are first defined. What follows are properties of norm space, grey relational space, grey relational grade for both globalized and localized grey relationships.

The GRA includes local relation and global relation analysis. Grey relational-based classifier is used to classify promoter/non-promoters. GRA is a method to determine the relation of a discrete data to other sequence data (Chang, 2000; Wu and Chen, 1999). The novel grey relational procedure is introduced. Suppose the test sequence φ_{t}(0), i=1,2,3,...,m, and K comparative sequences φ_{k}(0), φ_{k}(1),..., φ_{k}(m), i=1, 2, 3, ..., K, can be represented as

\[ \Phi_{\text{test}} = \{φ_{1}(0), φ_{2}(0), ..., φ_{l}(0) \} \]

\[ \Phi_{\text{com}} = \{φ_{1}(k), φ_{2}(k), ..., φ_{l}(k) \} \]

\[ \phi_{k}(k) = \{φ_{1}(k), φ_{2}(k), ..., φ_{l}(k) \} \]

Compute the absolute deviation of the test sequence \( \Phi_{\text{test}} \) and k comparative sequence \( \Phi(k) \) by

\[ Δφ_{l}(k) = [φ_{l}(0) - φ_{l}(k)] \]

The deviation matrix ΔΦ can be represented as

\[ Δ\Phi = \begin{bmatrix}
Δφ_{1}(1) & Δφ_{2}(1) & ... & Δφ_{l}(1) \\
Δφ_{1}(2) & Δφ_{2}(2) & ... & Δφ_{l}(2) \\
Δφ_{1}(3) & Δφ_{2}(3) & ... & Δφ_{l}(3) \\
... & ... & ... & ...
\end{bmatrix} \]

The grey relational grades r(k) can be calculated as:

\[ r(k) = \exp \left[ -ξ \left( \frac{ED(k)}{Δφ_{\max} - Δφ_{\min}} \right)^2 \right], \quad ξ \in (0, 25) \]

\[ = \exp \left[ -ξ \left( \frac{\sum_{i=1}^{m} (Δφ_{i}(k))}{Δφ_{\max} - Δφ_{\min}} \right)^2 \right] \]
\[
\Delta \varphi_{\text{max}} = \max_{\forall k} \max_{\forall l} \Delta \varphi_l(k)
\]

\[
\Delta \varphi_{\text{min}} = \min_{\forall k} \left[ \min_{\forall l} \Delta \varphi_l(k) \right]
\]

Where \( ED(k) \) is the Euclidean Distance(ED) between vector \( \Phi_{\text{ref}} \) and vector \( \Phi(k) \); \( \Delta \varphi_{\text{min}} \) and \( \Delta \varphi_{\text{max}} \) are the minimum and maximum values of the matrix \( \Delta \varphi \), respectively; \( \xi \) is a recognition coefficient with parameter interval \((0,25)\); \( \xi = 15 \) was chosen in this study. The grey relational grades \( r(k) \) are inversely proportional to the distances. If the test vector \( \Phi_{\text{ref}} \) is similar to any comparative vector \( \Phi(k) \), the grade \( r(k) \) will be a maximum value. GRA uses the grey relational grade to measure the relationship between the reference sequence data and comparative sequences data. The dimension of grey relational vector \( r=[r(1),r(2),...,r(k),...,r(K)] \) can be reduced from \( K \)-dimension to \( m \)-dimension by

\[
\gamma_j=\sum_{k=1}^{K} r(k)w_{kj}, \text{ } j=1,2,3,...,m
\]

\[
w_{kj}=\begin{cases} 
1, & k \text{ Class } j, \\
0, & k \notin \text{ Class } j.
\end{cases}
\]

The final grey grade \( g_j \) that an unknown vector \( \Phi_{\text{ref}} \) belongs to Class \( j \) can be derived from the following equation:

\[
g_j = \frac{\gamma_j}{\sum_{j=1}^{m} \gamma_j}, \text{ } j=1,2,3,...,m
\]

which defines the decision for classifying an unknown vector \( \Phi_{\text{ref}} \).

6. RESULTS AND DISCUSSION:

The following parameters: sensitivity (Sn), specificity (Sp), and Precision(Pr) are used to evaluate the predictive performance of the classifier. Let TP Number of true positive instances, FN Number of false negative instances, FP Number of false positive instances and TN Number of true negative instances.

\[
\text{True Positive Rate(TPR)/Sensitivity(Sn)/Recall} = \frac{TP}{(TP+FN)}
\]

\[
\text{False Negative Rate(FNR)/Miss} = \frac{FN}{(TP+FN)}
\]

\[
\text{True Negative Rate(TNR)/Specificity(Sp)} = \frac{TN}{(TN+FP)}
\]

A confusion matrix is drawn between the targets and outputs of the classifier.

The diagonal cells show the number of promoters/non-promoters that were correctly classified for each class. The off-diagonal cells show the number of promoters/non-promoters that were misclassified (e.g. promoters of Homo Sapiens were classified as Promoters of E.Coli etc.). The True Positive Rate/Sensitivity and False Negative Rate/Miss for each class are presented in the last row in green and red colors respectively. The Precision of each class is indicated in the last column(in green). The blue cell shows the total percentage of correctly predicted promoters/non-promoters (in green) and the total percentage of incorrectly predicted promoters/non-promoters(in red).

A Receiver Operating Characteristic (ROC) curve, a plot of the true positive rate (sensitivity) versus the false positive rate(1-specificity) is also drawn. As the curves are towards the Y-Axis and away from the X-Axis, the performance of the classifier can be considered as good.
7. CONCLUSION

The successful prediction of promoters with high accuracy using Grey Relational Analysis clearly indicates that the novel method has a promise as an approach for successful Prokaryotic and Eukaryotic promoter prediction. The experience gained from the above example shows that n-mer frequencies and Grey Relational Analysis is quite suitable to classify between promoter and non-promoter regions. The main aim of this paper is to develop an efficient tool that can discriminate between promoter and non-promoter in a given sequence with high accuracy. High result accuracy of the program indicated that the novel approach can be further successfully used for the prediction of Eukaryotic promoters in entire chromosome. This method is currently applied for estimating the number of promoters in different chromosomes of the human genome. Another challenge being addressed is the localization of promoters rather than a simple classification similar to the one at present. It is expected that the promising results using GRA will improve the performance of bio-molecular sequence analysis and promoter prediction in particular.

8. ACKNOWLEDGEMENTS

The authors would like to thank GITAM University, JNTUK, Kakinada and Acharya Nagarjuna University for providing computational facility and access to e-journals to carry out this research.

REFERENCES:


Stornoo, G. D. and D. Haussler (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. Int Conf Intell Syst Mol Biol. , Stanford, California, USA.


