SYNERGISTIC USE OF PROMOTER PREDICTION ALGORITHMS: A CHOICE FOR SMALL TRAINING DATASET?

By

Ekow CruickShank Oppon.

UNIVERSITY of the
WESTERN CAPE

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT

For the degree of

PHILOSOPHIAE DOCTOR

IN THE DEPARTMENT OF COMPUTER SCIENCE,
UNIVERSITY OF THE WESTERN CAPE.

Supervisor: Dr. Winston Hide
Co-supervisor Isabella Venter
December, 2000.

I declare that “**Synergistic use of Promoter Prediction algorithms: a choice for minimal training set**?” is my own work and all the sources I have quoted or used have been indicated and acknowledged by means of complete references.

Ekow Oppon’s Signature ……………………………………………….

Date ……………………………………………….

UNIVERSITY of the
WESTERN CAPE

ACKNOWLEDGEMENTS

My sincerest gratitude and appreciation goes to all the individuals and organizations that directly or indirectly assisted me in the completion of this research dissertation. My special thanks to:

→ Dr Winston Hide, my supervisor and mentor for his support, guidance, patience and direction that brought about the completion of this work;

→ The National Research Foundation (NRF), for sponsoring my studies since 1994 at both Msc. and Phd level;

→ South African National Bioinformatics Institute for providing the environment and facilities to enable me to complete the research. I am sincerely grateful to all the system administrations. Without them, the research would have taken a longer period.

→ The ElectroGenetics team especially Gary and Antoine for giving me valuable assistance and direction throughout the course of this work

→ My colleagues at South Africa National Bioinformatics Institute (Junaid, Alan and Andrey) who kept my morale up throughout the difficult times and offered their valuable assistance when needed.

→ Sandra Waterboer, for keeping the administrative worries out of my way so that I could put most of my concentration on the work and bringing up smiles during difficult times.

→ The Computer Science Department at University of the Western Cape and the late Prof. Stef Postma for registering at the department.

→ Peter Van Heusden and Wilson Honu for introducing the concept of abstract programming to me and filling all the ‘potholes’ at the initial stages of the course.

→ My dearest children Akwaaba, Euraffie and Akosua for involuntarily accepting to come second to this work in the four-year period used to accomplish this work.

UNIVERSITY of the
WESTERN CAPE

→ Weziwe, Simi and my family for their understanding and support throughout the period of the study.

TABLE OF CONTENTS

UNIVERSITY of the WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

TABLE **PAGE NUMBER**

UNIVERSITY of the WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

Results obtained by computing the dinucleotide composition of large data sets (+8000 per data) of promoters (P) and non-promoters (NP) of *E.coli, B.subtilis* and *Mycobacterium*. Promoter and Non-promoter data for both *E.coli* and *B.subtilis* consisted of 8000 nucleotides each whilst Mycobacterium promoter and non-promoter datasets constituted 5000 nucleotides each. Outstanding differences in composition of between promoters and non-promoters of certain dinucleotides are observed in all three organisms. They include TT, AA, AT and in *E.coli* and *B.subtilis,* and GC and CG in mycobacterium.

Table 5.3. 136

Percentage composition of all sixty-four triplets in promoter (P) and non-promoter (NP) of the three organisms namely *E.coli*, *B.subtilis* and *Mycobacterium*. Equal sizes (numbers and fragment sizes) of nucleotides in their natural genomic environment were analyzed. Triplets with difference of one percent or more (+1%) are highlighted in bold.

UNIVERSITY of the
WESTERN CAPE

Table 5.4. 141

False positive results obtained from the individual hash tables generated from promoter and non-promoter sequences of the same size (number of sequences and sequence lengths). Tested sequences have the same fragment sizes as the sets (promoter/non-promoter) used to develop the table. Five random sequences were generated from each of the original test sequences (101 bp) to obtain results very reflective on the actual test data.

Table 5.5. 142

The procedure used to obtain the data is similar to that used to obtain results in table 5.4. However, datasets have sequences of 75 bp fragment size each. The average numbers of false positives together with their respective percentage are shown in columns seven and eight.

Table 5.6. 143

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
B.	Bacillus
bp	base pair(s)
BCG	Bacillus Calmette Guerin
Bs	B.subtilis
cDNA	Complimentary DNA
DNA	Deoxyribonucleic acid
DFDA	Dinucleotide Frequency Distribution Analysis
Ec	E.coli
egg.	Example
et al.,	And others
FP	False Positive
fig.	Figure
html	Hyper text Marker language
HMM	Hidden Markov Model
kb	kilobase pair
M.	Mycobacterium
Mt	M.tuberculosis
mtub	M.tuberculosis
NN	Neural Network
NP	Non-Promoter
NNPP	Neural Network Promoter Prediction
No/no	Number
P	Promoter
Pos	Positive
RNA	Ribonucleic acid
SGI	Silicon Graphics International
TFDA	Triplet Frequency Distribution Analysis
TP	True Positive
UCE	Upstream Control Element

UNIVERSITY of the
WESTERN CAPE

INTRODUCTION

Recent exponential increases in DNA sequences due to advances in sequencing technology have brought with it new challenges such as new approaches to gene detection, promoter detection/prediction and homologous pattern recognition. Computational methods, as compared to traditional laboratory methods previously used in finding genes and protein binding sites, have therefore become unavoidable. Many biologists are rising to the occasion, coming up with hosts of algorithms that meet several of these new challenges. There are currently for example, many gene prediction packages that are accessible via the internet for both eukaryotes and prokaryotes. Among others, they include GeneMark (http://genemark.biology.gatech.edu/GeneMark), Orpheus (http://pedant.mips.biochem.mpg.de/orpheus), GeneID (http://kisac.cmb.ki.se/senn/internet/interne-geneid.html), GRAIL (http://avalon.epm.ornl.gov/GRAIL/), Genie (http://www-hgc.lbl.gov/proj-ects/genie.html), GENSCAN (http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html, HMMGENE (http://www.cbs.dtu.dk/services/HMMgene/), NetGene2 (http://www.cbs.dtu.dk/ser-vices/NetGene2) and GeneParser (http://beagle.colorado.edu/~eesnyder/GeneParser.html) among others. A list of these gene prediction programs together with the corresponding links to their websites can be found at the following url: http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-geneid.html. It is out of scope of this introduction to discuss various theories and algorithms behind these prediction packages. However, it is worth mentioning that, the methods of gene detection/prediction used by these programs cover statistical analytic and training/learning methods such as artificial neural network, Markov models, hidden Markov models and Bayesian networks. The varieties of algorithms/methods that are being applied to sequence analysis studies are perhaps an indication of the commitment that biological sequence analysts have put into annotating and elucidating the functional mechanism of genomes. Current genome annotations vis

UNIVERSITY of the
WESTERN CAPE

a vis gene predictions kept at Genbank are not complete unless annotations of the respective promoters of the corresponding genes promoter are carried out.

Promoter detection/prediction in a relatively difficult area of research. Prokaryotic promoter detection/prediction is probably less researched, if the current available number of promoter prediction tools on internet is used as a measure. Thus, whereas there are quite a number of promoter detection systems available on the internet (http://www.hgmp.mrc.ac.uk/GenomeWeb/nu-geneid.html) most of them have been developed for eukaryotes. Eukaryotic promoter detection systems currently available on the internet include Autogene (ftp.bionet.nsc.ru/pub/biology/aug), GeneID/Promoter2.0 (Knudsen, 1999), PromFind (Huchinson, 1996), PromoterScan (Prestridge, 1995); TSSG and TSSW (Solovyev and Salamov, 1997), PromoterInspector (http://genomatix.gsf.de/accounts/Help/PromoterInspector_help.html), NNPP (http://www-fruitfly.org/seq_tools.promoter.html). Also available is the Eukaryotic Promoter Database (EPD) at the url: http://cmgm.stanford.edu/help/manual/databases/epd.html#search. The prokaryotic prediction/detection on the internet, neural network promoter prediction (NNPP) was developed using artificial neural network system. A preliminary test of NNPP on a data set of five *E.coli* promoters and twenty-six (26) *E.coli* coding sequences of 75 bp sequence length using a threshold of 6.0 resulted in 3/5 (60%) true positives (TP), 2/5 (40%) false negatives (FN), 13/26 (~50%) false positives (FP) and 13/26 (50%) true negatives (TN). This preliminary analysis revealed the predictive accuracy of the NNPP to be low, having high false positive rate predictions. This observation has already been noted by other researchers such as Fickett (Fickett, 1998). One probable reason why NNPP does not do better is because, it is not designed for a particular prokaryotic organism. Due to the variability of the transcriptional machinery in prokaryotes as reflected on the availability of several known sigma factors, promoter prediction in prokaryotes has to be at least species specific for it to be very effective. Also, species specific programs would be more accurate if enough training datasets are available.

However, some prokaryotic promoters have been found to transcribe genes found in different species. For example, Mycobacterium heat shock promoters have been found to function in *E.coli* (Stover *et al*., 1991). However, there are reasons to believe that, most promoters are gene specific and in most cases only transcribe genes in their respective genomes.

There are some prokaryotic promoter datasets available upon request. Among others are, *B.subtilis* sigma *A* promoters (Helmann, 1995) and *E.coli* promoter sequences (Hannah and Margalit, 1993). However, they are not catalogued as databases and constitute mostly experimentally determined promoters. There are still quite a number promoter sequences available in the respective genomes of organisms whose entire genomes have been completely sequenced or about to be completely sequenced that need to be elucidated and analyzed.

There have been many attempts directed at predicting promoter sequences associated with respective genes, especially in *E.coli*. The quest has become even more pressing due to the availability of a number of completely sequenced prokaryotic genomes. Prokaryotic promoter prediction methods used to date include Statistical Analysis (Horton and Kanehisa, 1992; Oppon and Hide, 1998), Hidden Markov Models (Yada *et al*., 1996; Pedersen *et al*., 1996), Word and Pattern Matching Analysis (Pesole *et al*., 1992; Bourn and Babb, 1995), Artificial Neural Network (Pedersen and Engelbrecht, 1995; O'Neil, 1989; O'Neil, 1992; Mahadevan and Ghosh, 1994; Lukashin *et al*., 1989). Other methods that have been used in promoter prediction are given in algorithms using Expectation Maximization (Cardon and Stormo, 1991), in Rigorous Pattern Recognition Analysis (Galas *et al*., 1984) and in Cluster Analysis (Ozoline *et. al*., 1997). Most of the above approaches have had some degree of success with the task of promoter prediction, but they also predicted many sequences that were not known to have promoter activity according to existing data on *E.coli* promoters. Pederson *et al.* (1996) combined Artificial Neural Network (ANN) and Hidden Markov Model (HMM) in a move to increase the accuracy of promoter prediction. Hypothetically, combined algorithms (two) are expected to perform better than a single algorithm if the predictions are filtered through each method and perhaps three models/algorithms even better than two methods. It is in this context that this research is being undertaken.

OBJECTIVES

The project/research is designed to answer the following questions:

(a) Whether there is a minimum promoter dataset (for most or all prokaryotes) that is needed to effectively train prediction systems on so as to output predictions of high accuracy.

(b) Determine which section of prokaryotic region, if any, can be classified as ‘true’ promoter region.

(c) To investigate the possibility of integrating more than two prediction systems together so as to come up with a more effective promoter prediction tool.

(d) Use the integrated prediction systems to create a database of *E.coli, B.subtilis and M.tuberculosis* promoter sequences.

UNIVERSITY of the
WESTERN CAPE

The initial focus has been set on *M.tuberculosis* because of tuberculosis epidemic in the country (South Africa), especially in the Western Cape. With the regulatory regions of the various genes well categorized, researchers will be able to focus on genes and their promoters and use the information in their effort to find a solution to the tuberculosis problem. A database of *M.tuberculosis* predicted promoters will be established and eventually for other prokaryotic organisms too. Lastly, a prediction system that would require minimal number of prokaryotic promoter sequences for training will be created at the website of South African National Bioinformatics Institute. This prediction system, expected to have high degree of accuracy will be available to the world scientific community. The promoter prediction system will incorporate Artificial Neural Network (ANN), Hidden Markov Model (HMM), and a statistical approach based on analysis of triplet nucleotide composition of promoter and non-promoter sequences (Triplet Frequency Distribution Analysis – TFDA). ANN and HMM were selected based on their availability. TFDA used in generating values for specific triplets is a creation of the author of this dissertation.

Chapter One

Review on Prokaryotic and Eukaryotic Promoters

UNIVERSITY of the
WESTERN CAPE

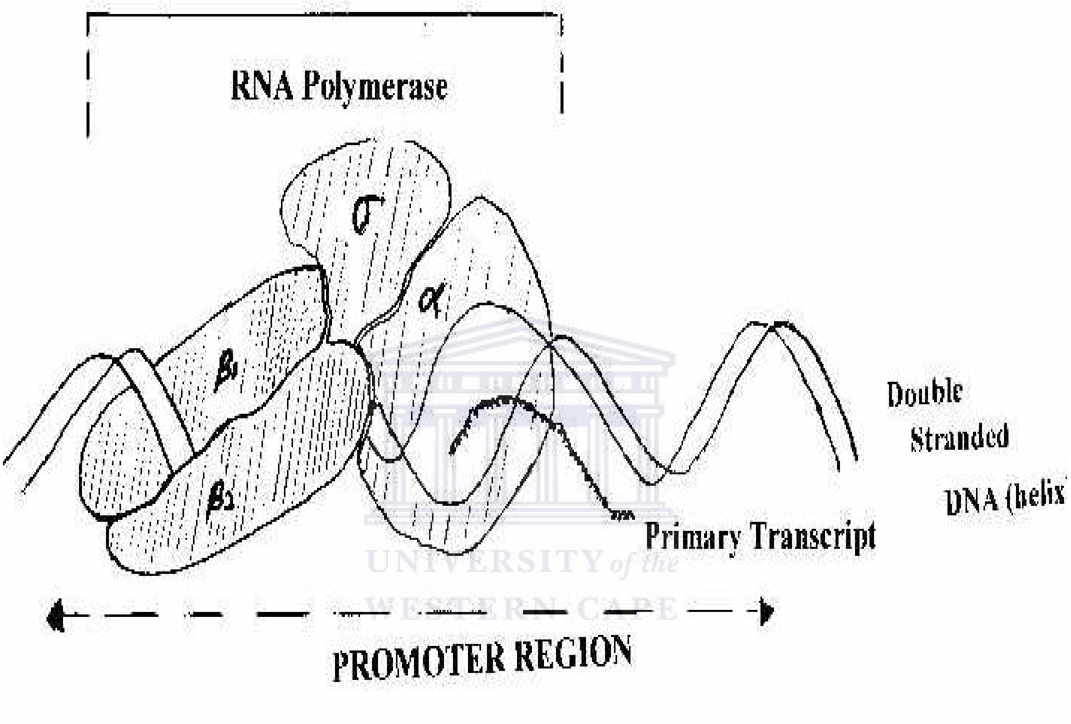
ABSTRACT

This chapter outlines basic gene structure and how gene structure is related to promoter structure in both prokaryotes and eukaryotes and their transcription machinery. An in-depth discussion is given on variations types of the promoters among both prokaryotes and eukaryotes and as well as among three prokaryotic organisms namely, *E.coli*, *B.subtilis* and Mycobacteria with emphasis on *M.tuberculosis*.

1.0. What is a Promoter?

The simplest definition that can be given for a promoter is: It is a segment of Deoxyribonucleic Acid (DNA) sequence located upstream of the 5’ end of the gene where the RNA Polymerase enzyme binds prior to transcription (synthesis of RNA chain representative of one strand of the duplex DNA). However, promoters are more complex than defined above. For example, not all sequences upstream of genes can function as promoters even though they may have features similar to some known promoters (from section 1.2). Promoters are therefore specific sections of DNA sequences that are also recognized by specific proteins and therefore differ from other sections of DNA sequences that are transcribed or translated. The information for directing RNA polymerase to the promoter has to be in section of DNA sequence defining the promoter region. Transcription in prokaryotes is initiated when the enzyme RNA polymerase forms a complex with sigma factors at the promoter site. Before transcription, RNA polymerase must form a tight complex with the sigma/transcription factor(s) (figure 1.1). The ‘tight complex’ is then converted into an ‘open complex’ by melting of a short region of DNA within the sequence involved in the complex formation. The final step in transcription initiation involves joining of first two nucleotides in a phosphodiester linkage (nascent RNA) followed by the release of sigma/transcription factors. RNA polymerase then continues with the transcription by making a transition from initiation to elongation of the nascent transcript.

WESTERN CAPE



UNIVERSITY of the WESTERN CAPE

Figure 1.1. Prokaryotic RNA polymerase comprising of the four subunits (α, β1, β2, and σ) in close complex formation with the nucleotide sequence of the promoter region. The above figure is not drawn to scale. The sigma (σ) unit of the enzyme is believed to be responsible for directing the RNA polymerase to the promoter.

A detailed study of promoters must encapsulate RNA polymerase together with transcriptional machinery. Thus, most discussions on promoters in this dissertation are given in context with RNA polymerase enzyme, transcriptional factors and processes involved in transcription. In studying prokaryotic promoter regions; the scope of current research, it is instructive to consider gene structure and the analogous, but more complex, eukaryotic promoter elements together with the associated transcriptional machinery.

UNIVERSITY of the
WESTERN CAPE

1.1. DNA and Gene Composition.

A gene in its simplest form may be defined as a sequence of deoxyribonucleic acid (DNA) containing codes for a gene product. DNA usually refers to polynucleotide strands twisted around each other in a double helix structure. Carbon phosphate backbones on the outside of the helix support nucleic acid bases adenine (A), thymine (T), guanine (G) and cytosine (C). Each polypeptide strand has a chemical polarity and is described as having opposite 5’ and 3’ ends. The polarity is based on the position of the carbon atom on the pentose ring to which phosphate groups bind in either direction. Any given region of the DNA helix might contain genetic information. The genetic code is read as a series of codons from the 5’ end of the gene that has the first nucleotide in the triplet codon. Each codon consists of three base pairs (bp) equivalent to a reading frame, which in turn corresponds to a single amino acid. There are 20 amino acids coded for by 61 triplet combinations from the four nucleotides. Genes may also be defined as the smallest functional unit of inherited genetic information that can be translated into a diffusible protein product or ribonucleic acid (RNA). Genes may be either housekeeping genes i.e. expressed in most tissues at all stages of development or tissue specific genes, i.e. requiring some degree of control over timing and levels of expression.

Genes may be divided into two classes; structural genes and regulatory genes. The products of structural genes are protein and RNA products. Regulatory genes as the name suggests code for protein products (structural genes) that are involved with the regulation of other genes. Regulatory genes together with cis-acting elements (sequence of DNA that functions only as DNA elements *in situ* affecting only the DNA to which it is physically linked) constitute control/regulatory elements. Cis-acting DNA elements include operators and repressors in prokaryotes, enhancers and silencers in eukaryotes, and promoters and terminators in both prokaryotes and eukaryotes.

WESTERN CAPE

1.2. Eukaryotic Promoters

Eukaryotic promoters consist of short sequence elements usually but not always found upstream of the transcriptional start site and are recognized by transcription binding factors. These *cis*-acting elements are usually spread out over a region of 200 bp. The function of the sequences between them are known to date, although the separation of the elements brought about the sequences that are not part of the element may have something to do with the conformation of the binding proteins. Some of these elements and the factors that recognize them are common: they are found in a variety of promoters and are used constitutively whilst other elements and their factors are specific to particular classes of genes (Lewin, 1997). The elements occur in different combinations in individual promoters. Accessory factors (transcription factors) are needed for transcription initiation but are not required for the subsequent elongation. The transcription factors other than RNA polymerase enzyme(s) are principally responsible for recognizing the cis-acting elements in the promoter region. Transcription initiation at an eukaryotic promoter therefore involves a large number of transcription factors that bind to a variety of cis-acting elements. An eukaryotic promoter may there be defined as the region containing all these binding sites. Thus the major feature defining the promoter for eukaryotic RNA polymerase is the location of binding sites for the transcription factors. Three types of RNA polymerase namely, RNA polymerase I, II, and III have been identified in eukaryotes. These three RNA polymerases bind to various kinds of promoters. Promoters used by RNA polymerases I and II are mostly upstream of the transcription start site. Some of the promoters used by RNA polymerase III are found downstream of the transcription start site. RNA polymerase I and III each recognize a relatively restricted set of promoters, and rely upon a smaller number of accessory factors (Reeder, 1984; Moss and Stefanovsky, 1995; Kahl *et al*., 2000). Accessory factors are proteins that help with transcription but do not make direct contact with the basal transcription factors. Promoters associated with the various RNA polymerases are discussed below.

UNIVERSITY of the
WESTERN CAPE

1.2.1. RNA polymerase I promoters.

Pre-ribosomal RNA is the sole transcript product RNA polymerase I. Consequently, RNA polymerase I requires recognition of only one kind signal in promoters for the expression of all genes it transcribes. RNA polymerase I has been found to be highly regulated to respond to both general metabolism (e.g. growth rate) and to specific environmental changes (Sollner-Webb and Tower, 1986; Reeder, 1990; Sollner-Webb and Mougey, 1991). Systematic analyses carried on the nucleotide sequences around the origins of transcription of some ribosomal DNA (rDNA) in different organisms revealed no common pattern among the nucleotide sequences (Sommerville, 1984; Moss *et al*., 1984) constituting promoters. The authors therefore suggested that, RNA polymerase I transcription system appear to have diverged considerably between organisms. They perceived the ribosomal transcription to be generally specific to taxonomic orders, the promoter of one group not being recognized by the transcription factors of another. Therefore, RNA polymerase I promoters have been thought to exhibit stringent (Grummt *et. al*., 1982) but not absolute (Pape *et al*., 1990) species specificity in its function. Some evidence suggests the existence of a common organization of all the promoters (Kownin *et. al*., 1985; Musters *et. al*., 1989; Firek *et. al*., 1990; Read *et. al*., 1992). These authors suggest that, the ribosomal promoter consists of essentially two domains or motifs. There is a `proximal promoter domain' (also called the minimal or core promoter) of ~45 bp, which includes the transcription start site. It is believed to be absolutely required for determining the accuracy of initiation. The other domain is an `upstream promoter domain' or 'upstream control element' (UCE), at about ~150 bp from the transcriptional start site.

Scientific literature reveal that, RNA polymerase I promoter has been best studied in human cells, where it has been found to consists of a bipartite sequence in the region preceding the transcriptional start site. The core promoter surrounds the start site extending from -45 to +20, and is believed to be sufficient for transcription to initiate. However, the efficiency of the core promoter (-45 to +20) is very much enhanced by the upstream control element (UCE), which extends from -180 to –107. Both the core promoter and the upstream control element have been found to have unusual composition for a promoter because they are very G•C-rich (Henderson and Sollner-Webb, 1990; Smith *et al*., 1993).

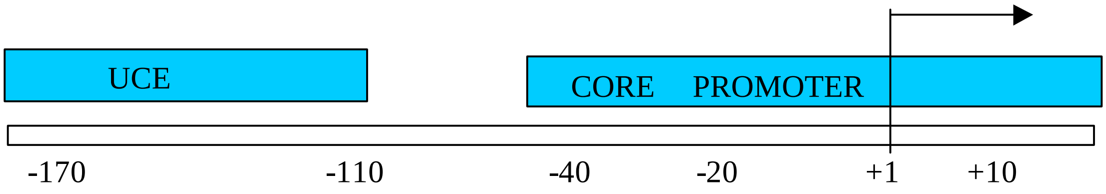


Fig.1.2. Eukaryotic RNA Polymerase I has a core promoter separated by ~70 bp from the upstream control element (UCE).

1.2.2. RNA polymerase II Promoters.

Promoters bound by RNA polymerase II are thought to be very diversified. Similarities of short sequences in the region near the start site are observed whenever promoters used by RNA polymerase II are compared. Analysis of mRNA transcripts around the start site revealed a high probability of first nucleotide of the start site to be A, flanked on either side by pyrimidines. This region around the start site has been defined as initiator, *Inr* (Smale and Baltimore, 1989) and it mostly consists of short weakly conserved motifs (Weis and Reinberg, 1992). The *Inr* is usually found between position -3 and +5. The transcriptional start site (tss) of RNA polymerase II promoters is usually identified by the *Inr* and/or by the TATA box close by. Mutational analyses have shown that, the initiator element is important for directing the synthesis of properly initiated transcripts (Goodrich *et al*., 1992) of all polymerase II promoters harboring *Inr*. The efficiency and specificity with which a promoter is recognized by RNA polymerase II however, is believed to depend on short sequences further upstream that are recognized by upstream, or inducible factors. These sequences and the factors that recognize them may be common for a wide variety of promoters, or they may be specific and particular for transcription in a restricted time or place (Nikolov and Burley, 1997). Three short sequences found around -30, -75 and -90 make up the core promoter. The TATA box (centered around -30) is the least effective component of the promoter as measured by the reduction in transcription that is caused by mutations (Lewin, 1997). Although initiation is not prevented when a TATA box is mutated, the start site of transcription varies from its usual precise location, confirming the

UNIVERSITY of the
WESTERN CAPE

role of the TATA box as a crucial positioning component of the core promoter (Wang and Stumph, 1995).

The sequence found around -75 is the CAAT box. It is often been found to be located up to –80. The CAAT box has been found to retain its promoter functionality at distances that vary considerably from the start site. Mutation experiments in and around the CAAT suggest that, the CAAT box plays a strong role in determining the rate at which the polymerase transcribes the adjacent gene(s). Though CAAT does not appear to play a direct role in promoter specificity, research has shown that, its increases promoter strength. Another element, the GC box, is found around -90 and usually contains the sequence GGGCGG. Multiple copies the GC box in either orientation are found in some polymerase II promoters. Promoters of RNA polymerase II appear are organized on a principle of "mix and match." Any combination of the promoter elements may contribute to promoter function, but none of the elements appear to be essential for all promoters.

UNIVERSITY of the
WESTERN CAPE

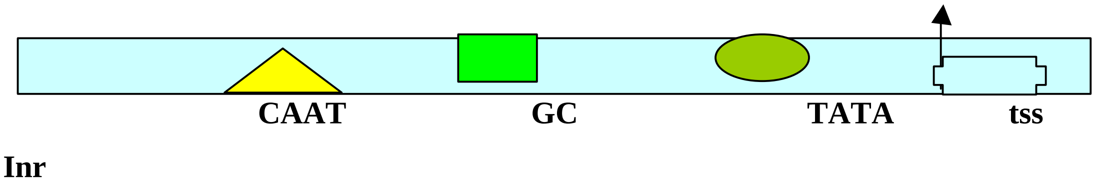


Figure 1.3. Cartoon depiction of RNA polymerase II promoter. Promoters are organized on a principle of `mix and match’. This means that, none of the elements is absolutely essential for promoter function but any combination of these elements is good enough for the RNA polymerase II to start transcription.

1.2.3. RNA Polymerase III promoters

UNIVERSITY of the
WESTERN CAPE

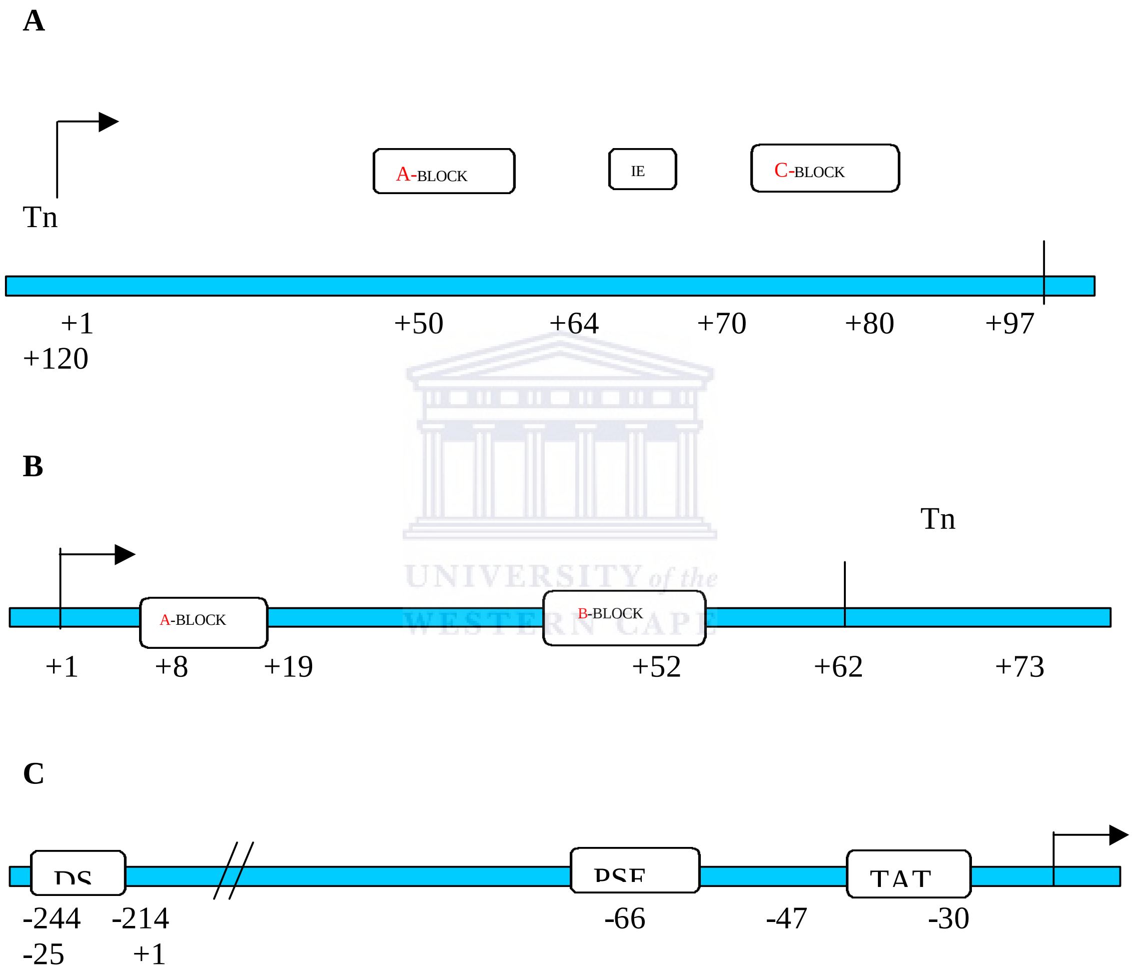
RNA polymerase III promoters can be categorized into two general classes recognized in different ways by different groups of transcription factors. The promoters for 5S and tRNA genes are described as internal; that is, they lie downstream of the transcription start site. Promoters for other genes such as small nuclear RNA (snRNA) genes are found upstream of transcription start site in the more conventional manner and belong to the second class of polymerase III promoters (Lobo and Hernandez, 1989; Tichelaar *et al*., 1994). The internal control regions required by class I RNA polymerase III promoters are generally composed of discontinuous elements of essential (necessary for promoter function) motifs separated by not yet functionally elucidated regions. An example can be found with the *Xenopus laevis* somatic 5S rRNA gene, which requires three internal elements for efficient transcription. These elements are: - an ‘A’ block, located between +50 and +64, an intermediate element (‘B’) at +67 to +72 and a C block from +80 to +97 (Roeder *et al*., 1987). The promoter is widely believed to relatively intolerant of changes in the spacing between individual elements (Roeder *et al*., 1987). The same type of internal region has also been found in the 5S rRNA genes of many lower organisms, including *D.melanogaster* (Sharp and Garcia, 1988) and *S.cerevisiae* (Lee *et al*., 1995). These promoters, described above are unique to 5S rRNA genes and are referred to as type I promoter (figure 1.4A).

The most common promoter arrangement in RNA polymerase III is found in tRNA genes of the adenovirus *VA* genes (Paule and White, 2000). Referred to as type II promoter, it is made up of two highly conserved sequence blocks named block A and block B within the transcribed region (Fig. 1.4B). The distance between block A and block B in a type 2 promoter have been found to vary quite extensively. Studies have revealed that, the boxes cannot often be brought too close together without abolishing promoter function (Fabrizio *et al*., 1987). The position of block B has been found to be extremely variable. Inter block separation of ~30-60 bp are said to be optimal for transcription, though a distance of around 365 bp from the start site have been found to be tolerated (Baker *et al*., 1987).

UNIVERSITY of the
WESTERN CAPE

A relatively minor group of polymerase III promoters have their promoters sequences located upstream of the transcriptional start site in the more conventional manner. These promoters have been grouped into the second class of RNA polymerase III promoters. Human and mouse U6 snRNA promoters are examples of promoters belonging to this group. The class two polymerase III promoters have been found to retain full promoter activity even after the deletion of all sequences downstream of transcriptional start sites (Lobo and Hermandez, 1989). Other promoters that have been found to have similar characteristics are human 7SK and MRP/7-2 RNA genes (Murphy *et al*., 1987; Yuan and Reddy, 1991).

The best characterized type III RNA polymerase III promoter belongs to a human U6 gene (Fig. 1.4C). The sequences required for efficient transcription are a TATA box, between –30 and –25, a proximal sequence element (PSE) between –66 and –47 and a distal sequence element (DSE) between –244 and –214 (Bark *et al*., 1987; Carbon *et al*., 1987).



UNIVERSITY of the
WESTERN CAPE

Figure 1.4. RNA polymerase III type I (A), type II (B) and type III (C) promoters. Type I promoter consists of bipartite sequences downstream of the start site, with boxA separated from boxB by intermediate elements (IE). Type II promoters (B) also consist of two boxes boxA and boxB found downstream of transcription start

site (+1). Type three promoters (C) consist of separated sequences upstream of the start site (DSE, PSE and TATA). Transcription termination sites are indicated by Tn.

1.3. Prokaryotic Promoters.

Prokaryotic promoters appear to be less complex (size and number of elements recognizable by sigma factors) than their eukaryotic counterparts though there are some similarities. For example, both are recognized by other factors before RNA polymerase binding. Prokaryotic promoters vary in their affinities for RNA polymerase, a factor very important with regard to controlling the frequency of transcription and therefore the extent of gene expression. Unregulated transcription initiation at many prokaryotic promoters have been found to require only an RNA polymerase holoenzyme, which consists of four core subunits with a dissociable σ factor. Multiple σ factors have been identified and each programs the core enzyme to transcribe from different class of promoters. Prokaryotic promoters direct not only the site of transcription initiation but also the rate of transcription. Earlier studies (Chamberlin, 1974; Hawley *et al*., 1982), have established that, promoter strength (as defined by degree which transcripts of the corresponding genes are produced) is primarily determined by two factors: the binding affinity to RNA polymerase and the rate of isomerization from ‘closed promoter complexes’ (DNA remains duplex) to ‘open promoter complexes’ (DNA opened by ‘melting’).

UNIVERSITY of the
WESTERN CAPE

Since the methods that will be used on mycobacterial promoter study will initially be applied on a study of *E.coli* and *B.subtilis* promoters, the promoters of these two organisms are also reviewed together with those of mycobacteria.

1.3.1. *E.coli* promoters

More than 300 promoters have been experimentally characterized by various researchers. A striking observation is lack of any extensive conservation of sequence over the 60 bp commonly associated with RNA polymerase interaction. There are four notable features in most *E.coli* promoters; the transcriptional start site, the -10 hexamer, the -35 hexamer and the distance between the -10 and -35 sequences. The transcriptional start site has been found to be purine in more than 90% of characterized promoters (Hawley and McClure, 1983). It is common for the transcription start site to be the central base within the sequence CAT, but the conservation of this triplet is not great enough to regard it as an obligatory signal (Rosenberg and Court, 1979; Siebenlist *et al*., 1980; Hawley and McClure, 1983). Just upstream of the start site, a six base pair (bp) region is recognizable in most promoters. The center of the hexamer is often close to 10 bp upstream of the tss. The distance varies in known promoters from 18 to 9 from transcriptional start site. Named for its location, the hexamer is often called -10 sequence. Its consensus is TATAAT and can be summarized in the form T_{80} $A_{95}T_{45}A_{60}A_{50}T_{96}$ where the subscripts denote the percent occurrence of the most frequently found base (figure 1.5). The other conserved hexamer is around ~35 bp upstream of the start site. The consensus for –35 has been universally accepted as TTGACA (Hawley and McClure, 1983). In more detailed form, the conservation is $T_{82}T_{84}G_{78}A_{65}C_{54}A_{45}$ (figure 1.5) (Hawley and McClure, 1983). The distance separating the -35 and -10 sites has been found to be between 16 and 18 bp in 90% of the promoters (Hawley and McClure, 1983). With very unusual exceptions, it may be as short as 15 bp or as wide as 21 bp. The distance may be critical in holding the two sites at the appropriate distance for the geometry of RNA polymerase (Olekhnovich and Kadner, 1999). An ideal *E.coli* promoter may consist of the -35 hexamer separated by 17 bp from the -10 hexamer with the –10 hexamer lying about 7 bp upstream of the start site. The -35 region is said to provide the signal for recognition by RNA polymerase, while the -10 sequence allows the complex to convert from `closed` to `open` form (Hawley *et al*., 1982).

Other researchers have established another important sequence element in addition to the four mentioned in some *E.coli* promoters (Newlands *et al*., 1992; Ross *et. al*.,

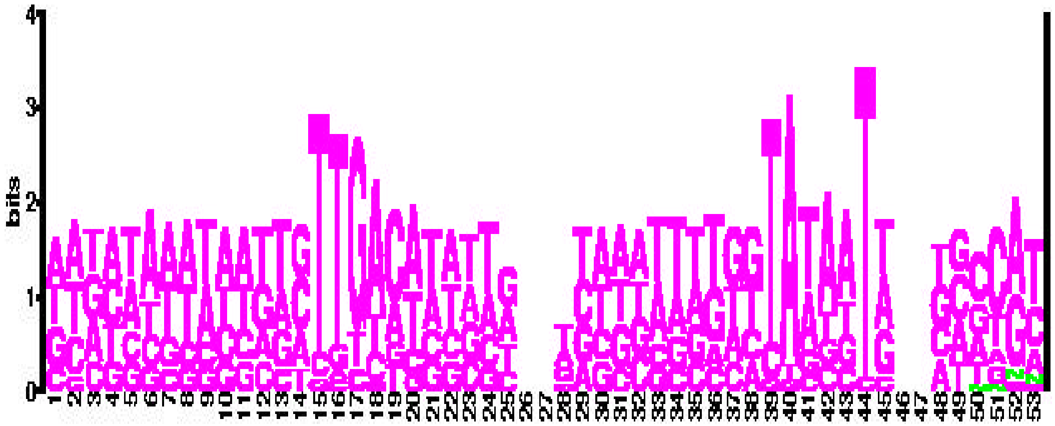
UNIVERSITY of the WESTERN CAPE

1993; Rao *et. al*., 1994). The seven *E.coli rrn* genes, which encode ribosomal RNA, are unusually strong, accounting for more than 60% of total RNA system in rapidly growing cells. The exceptional strength of the *rrn* promoter has been attributed to an AT-rich sequence of ~20 bp located immediately upstream of the -35 region. This region with the AT-rich motif has been termed upstream element or UP element (Ross *et. al*., 1993). The authors used two pieces of evidence to establish that UP element is recognized by RNA polymerase.

UNIVERSITY of the
WESTERN CAPE

Figure 1.5. Distribution of nucleotides around transcription start sites (position 51) of 115 *E.coli* promoter sequences. The canonical –35 (TTGACA) and –10 hexamers (TATAAT) are located at positions 15 to 21 and 39 to 44 respectively. Promoter data was obtained from Hawley and McClure (1983) and the informational analysis used is sequence logo (Schneider, 1997)

-35 and -10 hexamers



UNIVERSITY of the
WESTERN CAPE

First, the UP element was found to function in vitro in a transcription system containing only purified RNA polymerase and the promoter DNA sequences. The second evidence was in a DNAase I footprinting experiments, where RNA polymerase was found to protect the UP element yielding a ~20 bp extended footprint (Busby and Ebright, 1994). The UP element is believed to be functional as face of the helix phasing is maintained with respect to the transcriptional start site. The functional nature of UP elements when kept in phase with the helix was confirmed when mutations that change the spacer length in promoters altered the level of transcription in vitro (Ross et al., 1993). RNA polymerase has in general been found to tolerate changes in spacer length provided that they are compensated for by alterations in the conformation of the DNA, such as superhelix formation, so that the actual distance between the -35 and -10 signals remains the same (Ozoline and Tsyganov, 1995). The sequence immediately around the start site is believed to influence the initiation event and the initial transcribed region (from +1 to +30) influences the rate at which RNA polymerase clears the promoter and therefore has an effect upon promoter strength (Lewin, 1997). Thus, the overall strength of an

E.coli promoter cannot be predicted entirely from its -35 and -10 consensus sequences. A typical promoter may rely upon the -35 and -10 hexamers to be recognized by RNA polymerase, but one or other of these sequences can be absent from some exceptional promoters (Szoke *et al.*, 1987; Kobayashi *et al.*, 1990). In some of the cases, the promoters may not be recognized by RNA polymerase alone; it may require the intercession of ancillary proteins, which are thought to overcome the deficiency in intrinsic interaction between RNA polymerase and the promoter (Deuschle *et al.*, 1986; Keilty and Rosenberg, 1987; Belyaeva *et al.*, 1993).

1.3.2. *B.subtilis* Promoters

B.subtilis and *E.coli* promoters transcribed by either $E\sigma^{A}$ or $E\sigma^{70}$ have several similarities: the conserved sequences in the -35 and -10 hexamers, the distance between the two hexamers and the position of the transcription start site (Yamada *et al*., 1993). Thus most *B.subtilis* promoters normally function well in *E.coli* (Henkin and Sonenshein, 1987; Yamada *et al*., 1991; Chang *et al*., 1992). However, some functional *E.coli* promoters e.g. *lacUV5* are not transcribed by *B.subtilis* RNA polymerase (Henkin and Sonenshein, 1987). *B.subtilis* promoters have also been found to contain several moderately conserved sequences that may be the key to promoter being utilized effectively. These features may include A- and T-rich regions upstream of the -35 hexamer and *A* residues just downstream of the -10 hexamer (Helmann, 1995). In addition to these sequences, a region ending 1 base upstream from the -10 region appears to be conserved (Helmann, 1995). The sequence 5′-RTRTG -3′ (R = purine) was first found to be conserved in nine *B.subtilis* promoters and was termed the -16 region (Moran *et al*., 1982). A more comprehensive analysis of 142 promoters, all with experimentally determined transcription start site confirmed the conservation of the -16 region (Helmann, 1995). A ‘TG’ dinucleotide motif, positioned 1 base upstream of the -10 region was found in 45% of the *B.subtilis* promoters, T in 52% and the G in 58% of promoters (Helmann, 1995). The ‘T’ and the ‘R’ residues were also found to be correlated

UNIVERSITY of the
WESTERN CAPE

with the presence of the TG dinucleotide in some promoters (Helmann, 1995). Such promoters (extended –10) include a derivative of the *l Pre* promoter (Keilty and Rosenberg, 1987), the *galP1* promoter (Chan and Busby, 1989) and the *cysG* promoter (Belyaeva *et al*., 1993). The `extended -10 promoters' lack an identifiable -35 region but are transcribed by $E\sigma^{70}$ (Camacho and Salas, 1999). These promoters appear to bypass the need for a -35 region with the TG motif (Keilty and Rosenberg, 1987; Belyaeva *et al*., 1993; Chan *et al*., 1990). Point mutations in the TG motif of the *l Pre*, *galP1* and *cysG* promoters reduced or eliminated promoter function (Keilty, and Rosenberg, 1987; Chan, and Busby, 1989; Belyaeva *et. al*., 1993). The TG motif was found to reduce the temperature requirement for open complex formation by 20°C after being introduced into *galPcon6* promoter (Burns and Minchin, 1994). The reduction in temperature requirement may suggest that, the TG motif may be important in isomerization of promoter-enzyme-factor complex from a closed to an open complex in transcription initiation.

Further analysis of the dinucleotide composition of some more $E\sigma^{70}$ revealed A_2 (AA) and T_2-rich (TT) sequences in the upstream promoter region (-36 to -70) which are phased with the DNA helix: A_n tracts are common near -43, -54, and -65; whilst T_n tracts predominate at the intervening positions (Helmann, 1995). When compared with larger regions of the genome, upstream promoter regions have an excess of A_n and T_n sequences for n>4 (Helmann, 1995), where n denotes an integer. These data indicate that, an RNA polymerase binding site affects DNA sequence as far upstream as -70 (Helmann, 1995). Overall, the pattern of nucleotide conservation is reminiscent of that observed for *E.coli* promoters (Putzer and Leautey, 1994; Harley and Reynolds, 1987) and can be summarized as TTGaca (N_{17+-1}) TAtAAT (where bases in capital letters are present in more than 70% of promoters). As inferred from biochemical studies (Henkin and Sonenshein, 1987; Moran *et. al*., 1982), *B.subtilis* appears to be less tolerant of deviation from this 12 bp consensus than *E.coli*. On average, *B.subtilis* promoters match consensus at 9.1 positions compared with only 7.9 for *E.coli* (Lisser and Margalit, 1993; O'Neill, 1989). Perfect (12 out of 12) matches to this consensus are found in four out of the

UNIVERSITY of the WESTERN CAPE

125 chromosomal promoters (*glnR*, *rpmH*, *spoIIE* and *trnS*) but in none of 298 tabulated *E.coli* promoters (Lisser and Margalit, 1993). In addition, relatively few *B.subtilis* promoters (seven out of 125) lack an identifiable -35 region (less than 3/6 match to consensus), although not all of the assigned -35 regions are necessarily functional (Chassy and Murphy, 1993). Many other positions within the promoter have been found to exhibit a lesser degree of sequence conservation. Further statistical analysis revealed conservation of a T at -48, an A-rich region near -43, TnTG at -17 to -14 and a downstream extension of the -10 region (Helmann, 1995). Each of these features was noted previously based on an alignment of 29 promoters from several different gram-positive organisms (Graves and Rabinowitz; 1986), but they are not prominent in alignments of *E.coli* promoters (Harley and Reynolds, 1987; Hawley and McClure, 1983; Lisser and Margalit, 1993). The conserved -35 and -10 elements are most frequently separated by a 17 base spacer region as found for *E.coli* promoters (Helmann, 1995).

Many of the promoters used in *M.tuberculosis* studies were actually promoters from other mycobacteria. This is due to the unavailability of sufficient number of experimentally characterized *M.tuberculosis* promoters. As a result, *M.tuberculosis* promoters are discussed together with Mycobacterial promoters in general. Details on all the mycobacteria species and their corresponding promoters used in the study are documented on section 3.2.3.

WESTERN CAPE

1.3.3. Mycobacterial Promoters

1.3.3.1. Functionality in *E.coli*

Mycobacterial genomes have a high G+C content, for example, *M.tuberculosis* contains 65.9% G+C. Since the G+C content of a genome affects codon usage and promoter recognition (Nakayama *et al*., 1989; Ohama *et al*., 1987), it is expected that, transcription signals in mycobacteria may differ from those in other bacteria with different G+C composition such as *E.coli*. Although there are exceptions, mycobacterial promoters function poorly in *E.coli* (Sirokova *et al*., 1985; Das Gupta *et al*., 1993). Notable among the exceptions are mycobacteria heat shock promoters (Stover *et al*., 1991). Sequence similarities have been found between the mycobacterial heat shock promoters and consensus promoters recognized by σ70 and σ32 of *E.coli*. Among the mycobacterial promoters shown to be active in *E.coli* is the 16rRNA promoter of *M.bovis*. Suzuki *et al*. (1991), for example, expressed the *M.bovis* BCG 16S rRNA promoter *in vivo* and *in vitro* using the *E.coli* RNA polymerase. The authors identified a promoter upstream of the gene that showed similarity to *E.coli* promoters and was recognized by *E.coli* RNA polymerase. It was demonstrated that, the strengths of the *E.coli* and *M.bovis* BCG *rrn* promoters were identical when tested in *E.coli*. (Suzuki et al., 1991). The *E.coli* RNA polymerase did not however utilize another putative promoter of the BCG *rrn*, suggesting that, the second promoter may be recognized by a specific σ factor not present in *E.coli.* Other mycobacterial promoters that have been shown to function in *E.coli* are those associated with the 65 kDa antigens of *M.tuberculosis* (Shinnick, 1987), *M.bovis* BCG (Thole *et al*., 1987), *M.leprae* (Mehra *et al*., 1986) and the biotin carrier proteins of several species (Collins *et al*., 1987). More examples include *M.tuberculosis* 38 kDa antigen (Andersen *et al*., 1988), *M.paratuberculosis pAN* promoter clone (Murray *et al*., 1992), the *M. fortuitum blaF* (Timm *et al*., 1994), the *M.leprae* 18 kDa antigen (Dellagostin *et al*., 1995), the *M.tuberculosis katG* (Mulder, 1998) and promoter-containing clones isolated from *M. paratuberculosis* (Thomas *et al*., 1992). In all cases, expression in *E.coli* was less efficient than in the natural hosts (Mulder *et al*., 1997). Das Gupta *et al*.(1993), made libraries of *M.tuberculosis* H37Rv and *M.smegmatis* genomic DNA in an *E.coli*-mycobacterial shuttle vector containing a chloramphenicol acetyltransferase (CAT) reporter cassette and selected for clones expressing CAT. None of the *M.tuberculosis*-derived promoters and only 12 % of the *M.smegmatis*-derived promoter plasmids conferred chloramphenicol resistance on *E.coli* host cells. The

UNIVERSITY of the
WESTERN CAPE

authors suggested the existence of a good sequence similarity between *E.coli* and mycobacterial promoters at the -35 consensus, but significant variation at the -10 region. The authors used these promoters and other mycobacterial promoter sequences to generate the following probable consensus: -35: T (100 %), T (55 %), G (100 %), A (67 %), C (75 %), A (50 %); and -10: T (70 %), A (75 %), T (60 %), A (60 %), A/T (40 %), T (75 %). This study was however only limited to mycobacterial promoters that were known to be active in *E.coli.*

1.3.3.2. Promoters in both Fast and Slow growers (Mycobacterium).

Due to the slow growth and pathogenicity of *M.tuberculosis*, most of the promoters from this organism have been studied in either *M.smegmatis* or *M.bovis* Bacillus Calmette-Guerin (BCG) host. The expression of genes in fast growers such as *M.bovis* /M.smegmatis using promoters from slow growers, e.g. *M.tuberculosis* have provided evidence that, transcriptional signals are generally conserved among mycobacteria. Bashyam *et al*. (1996), for example, demonstrated that the efficiency and specificity of transcriptional recognition is conserved in *M.tuberculosis*, *M.smegmatis* and *M.bovis* BCG. The promoter clones examined in these three hosts exhibited similar activities and utilized the same transcription start sites. The authors suggested that *M.smegmatis* could be used as a surrogate host, at least for studying constitutively expressed *M.tuberculosis* genes. Similar results have been reported for the *M.tuberculosis* 16S rRNA (Verma *et al*., 1994), the *M.leprae* 18 and 28 kDa antigen and *M.bovis* BCG *hsp60* genes (Dellagostin *et al*., 1995). Although certain promoter sequences appears to be conserved among mycobacteria, there are likely to be differences in other aspects of the transcription machinery between the slow growers and the fast growers. The *M.smegmatis* transcription machinery has been shown to use the *M.bovis* BCG *hsp60* promoter in a similar manner to BCG. However, only one transcription start site was active in *M.smegmatis* (Levin and Hatfull, 1993). In addition, Timm *et al.* (1994), reported differences in the relative strengths of three mycobacterial promoters in

UNIVERSITY of the WESTERN CAPE

M.smegmatis and *M.bovis* BCG. Thus the viability of studying *M.tuberculosis* promoters in other mycobacterial hosts may depend on the particular promoter to be examined.

1.3.3.3. *M.tuberculosis* Promoters

Unlike *E.coli* and *B.subtilis*, relatively fewer *M.tuberculosis* promoters (~35 to date) have been experimentally characterized. and even less (~32 promoters) have their transcriptional start site experimentally characterized. However, many researchers have been actively involved in elucidating features characteristic of *M.tuberculosis* promoters. Kremer *et al*. (1995), carried out a detailed study of the promoter region of the *M.tuberculosis* 85A antigen gene. They made progressive deletions of the 5' end using nuclease *Bal31*. All of the deletions resulted in lower levels of expression than the full length fragment. Removal of the first 44 bp resulted in a 40 % decrease in promoter activity. Further studies revealed that, the essential promoter region to be between nucleotide -26 and -136 with respect to the translation initiation codon. The transcriptional start site was found to be located 63 bp upstream of the proposed ATG initiation codon. The -10 hexamer showed some similarities to other mycobacterial promoters and to some *Streptomyces* promoters (which were not expressed in *E.coli*). Two putative -35 regions were identified. One (17 bp from the -10 hexamer) showed 50% sequence similarity with that of σ^{70} promoters. The other (located 22 bp from the -10 region), showed 83 % identity with the *E.coli* σ^{70} consensus sequence and was identical to the -35 region of the *M.leprae* and *M.tuberculosis* 16S rDNA promoter regions.

Das Gupta *et al.* (1993) also isolated a number of *M.tuberculosis* H37Rv and *M.smegmatis* DNA fragments able to promote expression of the CAT reporter gene in *M.smegmatis*. They found that, the frequency of isolation of promoter clones was 10-20% for *M.smegmatis* (350 altogether) and 1-2 % for *M.tuberculosis* (125 altogether). Most of the promoters from *M.tuberculosis* gave CAT activities of 5-100 nmol/min/mg protein, while most of the *M.smegmatis* promoters gave much higher activity (>500 nmol/min/mg protein). The authors suggest that, strong

UNIVERSITY of the
WESTERN CAPE

promoters occur less frequently in *M.tuberculosis* than in *M.smegmatis*. This is consistent with the lower frequency of isolation of promoters from *M.tuberculosis* by Das Gupta *et al*. (1993). However, the observations may have been due to the expression of the *M.tuberculosis* promoters in a heterologous host (Mulder *et al*., 1997). Bashyam *et al.* (1996) sequenced 10 of the *M.tuberculosis* promoters isolated in the above-mentioned study and aligned them on the basis of their transcription start sites. All contained a conserved -10 region at similar positions upstream of their transcription start sites. The conserved sequences were T (80 %), A (90 %), Y (60 %), g (40 %), A (60 %), and T (100 %) where Y denotes a pyrimidine base.

As in *E.coli*, the first, second, and sixth nucleotides of the –10 region are most strongly conserved. The less conserved bases tend more towards G and C substitutions. None of the -35 regions of the promoters studied were homologous to the *E.coli* consensus sequence and none were conserved in the mycobacteria. The authors suggest that the absence of a conserved -35 region is a distinctive feature of mycobacterial promoters (Bashyam *et al*., 1996). This suggestion from Bashyam *et al*., (1996) is in contrast to the findings of Ramesh and Gopinathan (1995), but is supported by the results of Sarkis *et al*. (1995), Kremer *et al*. (1995) and Kenney and Churchward (1996). In other studies, deletion analysis of one *M.tuberculosis* promoter revealed the -35 region alone to be insufficient to support transcription and -10 region to be essential for transcription. In 9 of 14 *M.smegmatis* and 7 of 10 *M.tuberculosis* promoters, transcription initiated at a purine (Bashyam and Tyagi, 1998). *M.tuberculosis* promoters have a higher G + C content (57 %) from positions -1 to -50, with respect to the translation initiation codon, than the *M.smegmatis* promoters (43 %), which may have had a bearing on the lower strength of the *M.tuberculosis* promoters (Bashyam *et al*., 1996). Further support for the importance of the -10 region in promoter efficiency in the mycobacteria is provided by the isolation of up-mutations in promoter sequences. Point mutations in the upstream region, which result in overexpression, have been identified for *M.tuberculosis ahpC* genes (Dhandayuthapani *et al*., 1996; Sherman *et al*., 1996; Wilson and Collins, 1996; Heym *et al*., 1997).

UNIVERSITY of the
WESTERN CAPE

Although -10 and -35 hexamers play an important role in promoter function, other regions of the DNA upstream of genes can play a supplementary role. It has been found that, maximal expression of the *M.tuberculosis katG* promoter requires a 155 bp region 300 bp upstream of the translation start codon and approximately 200 bp upstream of the putative -35 region (Mulder, 1999). This ‘upstream activator region’ binds to one or more *M.smegmatis* proteins and contains a 24 bp AT-rich (66.67 %) sequence which is 79.2 % homologous to a region located 489 bp upstream of the *M. fortuitum katG* gene). These regions may be analagous to the AT-rich upstream (UP) elements found in *E.coli* that increase promoter activity (Ross *et al*., 1993). The presence of such a region upstream of the *M.tuberculosis katG* genes suggests common mechanisms of regulation between the *M.tuberculosis* and *E.coli katG* genes (Mulder, 1999). Another, possibly analagous region is a 41 bp sequence located 269 bp upstream of the *M.tuberculosis recA* gene which was found to be essential for expression. This region contains no functional promoters and may act as an upstream regulatory region by binding to an activator protein (Movahedzadeh *et al*., 1997).

UNIVERSITY of the
WESTERN CAPE

1.4. Is there a common structure for Promoters?

The interaction between protein and nucleic acids is an ancient and fundamental feature of evolution. Such interactions no doubt have under so many constraints through evolution. It is therefore expected that, sections of sequences that direct transcription of genes have had their ‘own’ kind of evolution, no doubt, orchestrated by the very genes they transcribe. Organisms have had to develop a system through evolution, where the ‘right’ genes had to be transcribed at the ‘appropriate’ times. The adoption and use of transcription factors has no doubt been a very successful strategy to the problem (transcribing vital genes when most needed). This use of different sigma factors to facilitate transcription has probably been made possible due to DNA-binding proteins in most cases acting at different sites where they display different activities. Such acts of DNA-binding proteins would also ensure large number of potential subtypes of binding sites for any

DNA-binding protein, probably explaining why certain promoter sequences of one organism function successfully in other organisms.

Thus, evolutionary requirements have necessitated the need by organisms to save resources and utilize them effectively, that is, transcribing genes whose products are needed. An apparent solution to the problem of efficient utilization of resources seems to be the use of sigma/transcription factors. The appropriate sigma/transcription factors are used to assemble the transcription machinery at the promoter region of the gene(s) to be transcribed. The signal for the positioning of the factors-enzyme complex, that is recognition of the promoter region therefore has to come from the sequence that define the promoters and probably the adjacent gene(s). However, as noted above, not all promoters appear to have the ‘known signals’ that are responsible for assembling factors and polymerase enzyme necessary for the transcription of the adjacent gene(s). Somehow, these very promoter regions are recognized by the respective factors and RNA polymerase enzyme. The problem of what is recognizable as the ‘signal’ is compounded by the fact that there are other regulatory sequences such as oppressors and operators that sometimes play major roles in transcription. A prerequisite for promoter function in both prokaryotes and eukaryotes appears to be, an AT-rich region that facilitates the opening of the DNA helix structure before transcription. That technically puts any AT-rich region in a genome as a potential promoter region, but does not necessarily make every AT-rich region a DNA-binding site. Perhaps, the driving force behind the recognition of any promoter region is the adjacent gene(s) to be transcribed since similar or ‘stronger’ promoter-like sequences have not been to have promoter function (personal observation). In any case, certain sequences have features that mask them as promoter sequences. Though not a perfect system to the human mind as no common feature has not been observed for all promoters, these sequences exist and they are recognized by the factors and the enzymes that need to recognize them. If these promoter sequences can be recognized by the various factors and the polymerase enzyme, then it is possible for methods/systems to be developed that will recognize them too. As to whether there is a common promoter

UNIVERSITY of the
WESTERN CAPE

structure, perhaps there much more to be learnt that would change the way we perceive structural organization in living cells.

UNIVERSITY of the
WESTERN CAPE

Chapter two.

Hidden Markov Model, Artificial Neural Network and Triplet Frequency Distribution Analysis.

ABSTRACT

Three algorithmic approaches: Hidden Markov Model (HMM), Artificial Neural Network (ANN) and Triplet Frequency Distribution Analysis (TFDA) have been selected to be used for study. The study is on the ability to of the three methods to learn and predict promoter sequences from non-promoter sequences. The prediction systems (HMM, ANN and TFDA) will be exhaustively assessed independently on known promoter sequences of the three organisms The three prediction systems will then be combined and used in predicting promoter sequences from entire genomes of *E.coli, B.subtilis* and *M.tuberculosis*. In this chapter, brief introductions are given on HMM and ANN whilst the rationale, principle and theory behind TFDA is reviewed.

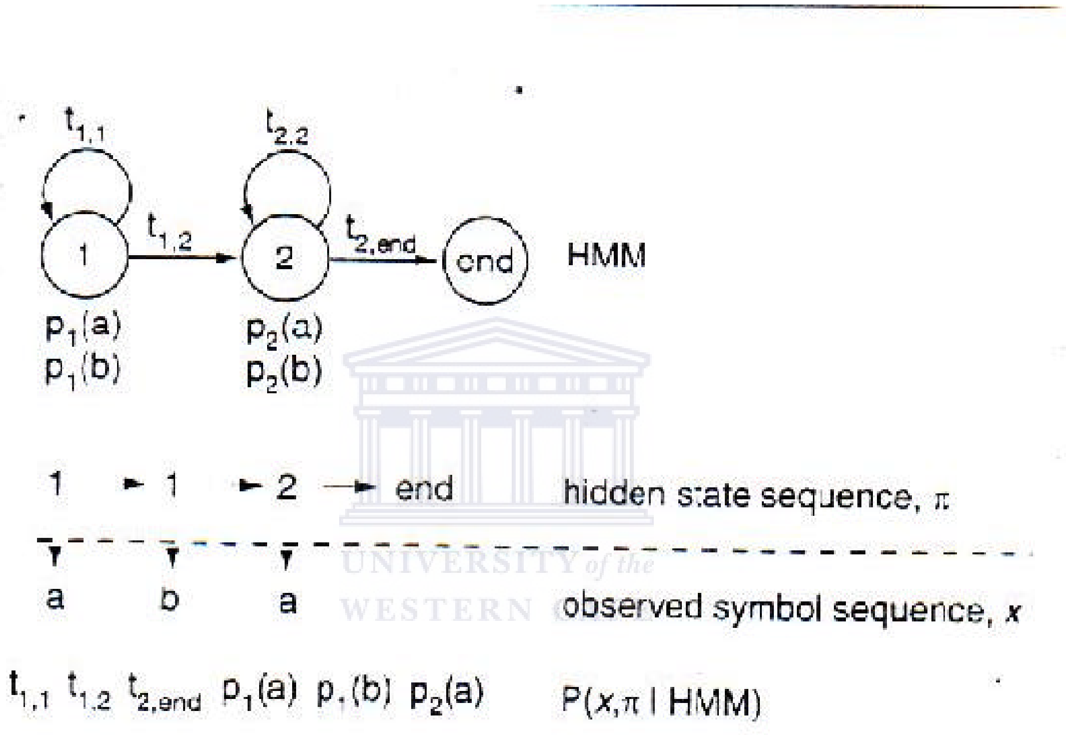
UNIVERSITY of the
WESTERN CAPE

2.1. Hidden Markov Models.

2.1.1. Introduction

A HMM describes a probability distribution over a certain number of sequences. Because a probability distribution must sum to one, the ‘scores’ that a HMM assigns to sequences are constrained within 0 and 1. Thus the increase in probability of one sequence will result in a decrease in the probability of one or more other sequences. An simple HMM that models sequences of two letters (*a*,*b*) is shown in figure 2.1.1. The modeled HMM illustrates a problem in which sequences started with one residue composition (*a*-rich), then switched once to a different residue composition (b-rich). The HMM consists of two states connected by state transitions. Each state has a symbol emission probability distribution for generating state transitions. Each state has a probability distribution according to whether it matches a specific symbol in the sequence. Starting in an initial state, a new state with some transition probability is selected. This new state may be 1, with transition probability *t1,1*, or state 2 with transition probability *t1,2*. Then a residue with an emission probability specific to that state is generated. The transition/emission continues until the end where an end state *s*. At the end of the process, there is hidden state sequence that is not observed and a symbol sequence that is observed. The name `hidden Markov model' comes from the fact that the state sequence is a first-order Markov chain, but only the symbol sequence is directly observed.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 2.1.1. A HMM modeling sequences of *as* and *bs* as two regions of potentially different residue composition. Circles represent states whilst arrows represent state transitions. A possible state sequence generated from the model is shown, followed by a possible symbol sequence. The joint probability of the symbol sequence and the state sequence is given by the product of all transition and

emission probabilities. In HMM, the state sequence (e.g. the biologically meaningful alignment) is not uniquely determined by the observed symbol sequence, but must be inferred probabilistically from it. Diagram copied from Sean Eddy's publication entitled 'Profile hidden Markov models (Eddy, 1998).

HMM states may be associated with meaningful biological sequences such as the position(s) of certain nucleotides in a motif. In the above described HMM for instance, states 1 and 2 may correspond to a biological notion of two sequence regions with different residue composition. Inferring the alignment of the observed protein or DNA sequence to the hidden state sequence is like labeling the sequence with relevant biological information (Barett *et al*., 1997).

An HMM can be built from a set of unaligned sequences by iteratively estimating the transition/emission probability parameters from the sequence as various alignment options are considered. Alternatively, a HMM can be built from pre-aligned sequences, that is where the state paths are known. In the latter case, the parameter estimation problem is simply a matter of converting observed counts of symbol emissions and state transitions into probabilities.

Standard HMM training algorithms include Baum-Welch expectation maximization or gradient descent algorithms. Simulated annealing and genetic algorithm training methods have been found to be better at avoiding spurious local optima in training HMMs and HMM-like models (Eddy, 1996; Neuwald *et al*., 1997; Durbin and Holmes, 1998). Most training algorithms seek relatively simple maximum likelihood (or maximum a posteriori) optimization targets. More sophisticated optimization targets are used to compensate for non-independence of example sequences e.g. biased representation (Eddy, 1996; Bruno, 1996; Durbin and Holmes, 1998; Karchin and Hughey, 1998; Sunyaev *et al*., 1998), or to maximize the ability of a model to discriminate a set of true positive example sequences from a set of true negative training examples (Mamitsuka, 1996).

UNIVERSITY of the
WESTERN CAPE

Proteins, RNA and other features, including promoters in genomic DNA sequence can be classified into families of related sequences and structures (Hennikoff *et al*., 1997). Multiple alignments of a sequence family reveal relatedness in their pattern of conservation. Some positions are more conserved than others, e.g. the –35 and –10 boxes of *E.coli* promoter sequences, while some regions of a multiple alignment appear to tolerate insertions and deletions more than other regions. Thus, position specific information needs to be incorporated in algorithms and models used in database searches for similar sequences. HMMs (Haussler *et al*., 1993; Krogh *et al*., 1994) and related generalized profiles (Bairoch and Bucher, 1994) have been used with some degree of success in detecting conserved patterns in multiple sequences (Baldi *et al*., 1994; Eddy, 1995; Eddy *et al*., 1995; Bucher *et al*., 1996; Hughey and Krogh, 1996; McClure *et al*., 1996; Eddy, 1998). HMMs are useful as formal fully probabilistic forms of profiles (Baldi *et al*., 1994; Eddy *et al*., 1995; Krogh *et al*., 1994; Stultz *et al*., 1993). They wield a mathematically consistent description of insertions and deletions and also offer theoretical insight into the difficulties of combining disparate forms of information such as in sequences (Eddy, 1994). One of the features of HMMs is that it is possible to train models from initially unaligned sequences, thus producing HMM-based multiple alignments (Baldi *et al*., 1994; Krogh *et al*., 1994). HMMs can therefore be used to build ‘profiles’ of promoter sequences that can be used in database searches for other uncharacterized promoter sequences.

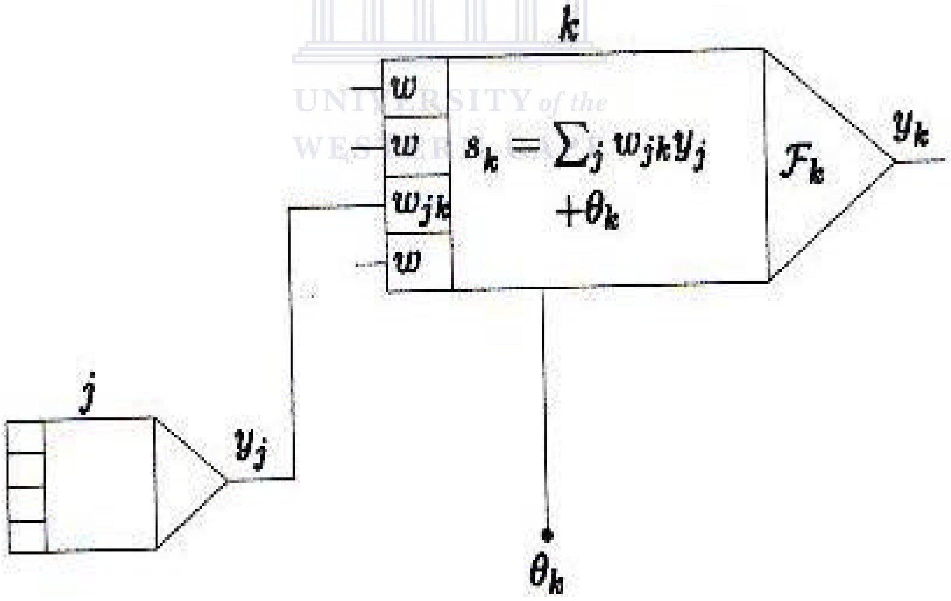
2.2. Artificial Neural Network.

Most of the literature presented below on artificial neural network together with those on the various network topologies was obtained from various websites on the internet. They constitute mainly lecture notes and slides from academic institutions. The websites include: http://www.cs.nott.ac.uk/~sbx/winnie/aim/neural, http://www-dse.doc.ic.uk/~nd/surprisek/ jour-nal/vol14/cs11/report.html#Human, http://www.interstate95.com/home/adaptive.

UNIVERSITY of the
WESTERN CAPE

2.2.1 Introduction

Artificial neural networks can be most adequately characterized as 'computational models' modeled on biological neurons with peculiar properties, such as ability to adapt or learn, to generalize and to cluster or organize data. It is an information processing system made up a number of very simple and highly interconnected processors called neurons. The most important aspect of neural net architectures is the fact that they consist of these simple and highly interconnected processors, the neurons. Generally, nodes within all neural networks follow a common model of operation. They sum their input signals, pass the summed value through an activation function and send that value out as its output signal. The output signal will either leave the network or will 'travel' along a connection to another node and act as input to that node. These neurodes are the analogs of the biological neural cells, or neurons in the brain. There are two primary methods of training a designed neural network. Supervised training is akin to teaching a child by example. The neural net gets input signals presented at its input signal and corresponding correct output signals, and the network tries adjust it tunable parameters to capture the relationship between the input and the output. The second method, self-organization or unsupervised training allows the neural network to separate a set of training input patterns into various categories based on similarities and differences between the input signals.



UNIVERSITY of the
WESTERN CAPE

Fig. 2.2.1. The basic components of an artificial neural network. The propagation rule used here is the standard ’weighted’ summation. The total input to unit *k* is the ‘weighted’ sum of the separate outputs from each of the connected units (e.g. y_j) plus a *bias* or *offset* term θ_k. Unit k then passes on the `weighted’ summation as an input to another node (neuron) or as an output signal. The figure was obtained via internet from lecture notes on neural network at the Computer Science Department at Sheffield university.

2.2.2. Architecture

A major aspect of a parallel-distributed model of artificial neural network can be distinguished (McClelland and Rumelhart, 1986). It consists of the following features:

1. A set of processing units (‘neurons’ cells);
2. A state y_j of activation for every unit, which is equivalent to the output of the unit;
3. Connections between units. Generally each connection is characterized by a weight w_{jk}
   which determines the effect, which the signal of unit *j* has on unit *k*.
4. A propagation rule, which determines the effective input s_k of a unit from its external
   inputs;
5. An activation function $T_{k,}$ which determines the new level of activation based on the
   effective input $s_k(t)$ and the current activation $y_k(t)$ at period *t*;
6. An external input (aka bias, offset) q_k for each unit;

UNIVERSITY of the
WESTERN CAPE

7. A method for information gathering (the learning rule);
8. An environment within which the system must operate, providing input signals and if necessary, error signals;

Figure 2.2.1 illustrates these aspects of the architectural structures mentioned above. Each neural unit performs a relatively simple job; receive input from neighbors or external sources and use this to compute an output signal, which is propagated to other units or to network output. Apart from this processing, a second task during training is the adjustment of the ‘weights’. Neural network systems are inherently parallel in the sense that, many units can carries out their computation simultaneously and independently. Three types of units are identifiable. Input units (indicated by an index *i*), which receive data from the neural network environment. Output units (indicated by an index *o*), which send data to the neural network and hidden units (indicated by an index *h*) whose input and output signals remain within the neural network. During training, units can be updated either synchronously or asynchronously. With synchronous updating, all units update their activation simultaneously, whereas with asynchronous updating, each unit has a (usually fixed) probability of updating its activation at a time *t* and usually only one unit will be able to do this at a time.

UNIVERSITY of the WESTERN CAPE

2.2.3. Network Topologies

There are two major network topologies. Feed-forward networks and Recurrent networks.

2.2.3.1. Feed Forward Networks

Feed-forward networks: where the data flow from input to output is strictly feed-forward. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from output of units to inputs of units in the same layer. Classical examples of feed-forward networks are McCulloch-Pitts Neuron, the Perceptron and Adaline networks.

2.2.3.2. Recurrent networks

Recurrent networks do contain feedback connections unlike feed-forward networks. During training of recurrent networks, the activation values of the units at neurons undergo a relaxation processes such that, the network evolves to a stable state in which these activations do not change anymore. In other applications of recurrent networks, the change of the activation values of the output neurons are significant, such that the dynamical changes in values constitute the output of the network (Pearlmutter, 1990). Examples of recurrent networks include Kohonen (Kohonen, 1977) and Hopfield (Hopfield, 1982) networks.

UNIVERSITY of the
WESTERN CAPE

2.2.4. Training of artificial neural networks

A neural network has to be configured such that the application of a set of inputs produces (either ‘direct’ or via a relaxation process) the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using *a priori* knowledge. Another way is to ‘train’ the neural net by feeding it teaching patterns and letting it change its weights according to some learning rule.

2.2.5. Paradigms of learning

Learning situations can be categorized in two distinct types. Supervised learning, also referred to as Associative learning; in which the network is trained by providing it with input and matching output patterns. Unsupervised learning or Self-organization in which an (output) unit is trained to respond to clusters of pattern within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no *a priori set* of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli. When using neural network one has to distinguish two issues that influence the performance of the system. The first one is the representation power of the network; the second is the learning algorithm. The representational power of a neural network refers to the ability of a neural network to represent a desired function. Since neural networks are built from sets of standard functions, in most cases the network will only approximate the desired function. Even when the network has an optimal set of weights, the approximation error is never zero. The second issue is the learning algorithm. If an assumption is made that, there exists a set of optimal weights and these weights can be achieved, is there a procedure to iteratively find this set of weights? If these optimal weights can be achieved, the time duration that it takes to achieve the optimal weights must also be put into consideration.

UNIVERSITY of the
WESTERN CAPE

2.3. Triplet Frequency Distribution Analysis (TFDA)

2.3.1. Introduction

One of the many observations revealed by biological sequence analysis is the difference in DNA composition of coding and non-coding regions in genomes. Irrespective of the GC content of the organism in question, the differences in nucleotide content of the coding regions and non-coding/regulatory regions have been observed in many organisms. Thus, most bacterial gene prediction algorithms such as GeneMarkHmm (Besemer and Borodovsky, 1999) and Orpheus (Frishman *et al*., 1998) utilize the codon usage of the bacteria and the statistical differences of the nucleotide composition between coding and non-coding sections of the genome. The presence of the –10 and –35 consensus regions in some prokaryotic promoters including those *E.coli*, *B.subtilis* and *Streptomyces* also confirm that certain DNA arrangements are peculiar to promoters and/or regulatory sequences as compared to other regions in the genome. Attempts to utilize this information to study and conduct statistical-related analysis of nucleotide composition in DNA include; the correlation between the nearest neighbor bases (Josse *et al*., 1961; Gatlin, 1966) and the heterogeneity of base density in fragmented DNAs (Sueoka, 1959). Statistical regularities have also been used to detect coding regions (Shulman *et al*; 1981; Shepherd, 1981a; 1981b; Staden and McLachlan, 1982; Fickett, 1982; Frishman *et al*., 1998; Borodosky and Besemer, 1999). The same ideas and principles have been used to study nucleosome formation (Trifonov and Sussman, 1980) and promoter detection/prediction (Horton and Kanehisa, 1992; Oppon and Hide, 1998). These studies focused on particular aspects of the correlation structure of DNA sequences in relation to particular biological problems.

UNIVERSITY of the
WESTERN CAPE

2.3.2. Analysis of Nucleotide (Triplets) composition in DNA sequences.

TFDA is a statistical approach to promoter detection/prediction based on analyzing the information content of promoters and non-promoter sequences in the form of triplets (not codons). A unique hash table is generated for each promoter non-promoter set pair. The outline of the hash tables (relative frequency of nucleotide composition) are similar with respect to the relative values of the triplets. For each promoter non-promoter set pair, the frequency of each of the sixty-four (64) possible triplets (as a 3 bp window is shifted 1 bp along the sequence) is calculated for both pairs. The frequency value of each triplet from the non-promoter (f_{np}) then subtracted from the corresponding triplet frequency value in the promoter set (f_{p}) to generate a hash table of triplet differences (f_{p}-f_{np}), figure 2.3.1.

```
ACGTGCACATGCGTAACCGTGCATGCGTACGTACGATACAGTGCACTGA
ACG
 CGT
  GTC
   TGC
    GCA
     CAC
```

UNIVERSITY of the
WESTERN CAPE

Figure 2.3.1. An illustration of how triplets were obtained from sequences.

```
TGG =  0.5430 GGT = -0.1164 TAT =  0.1939
```

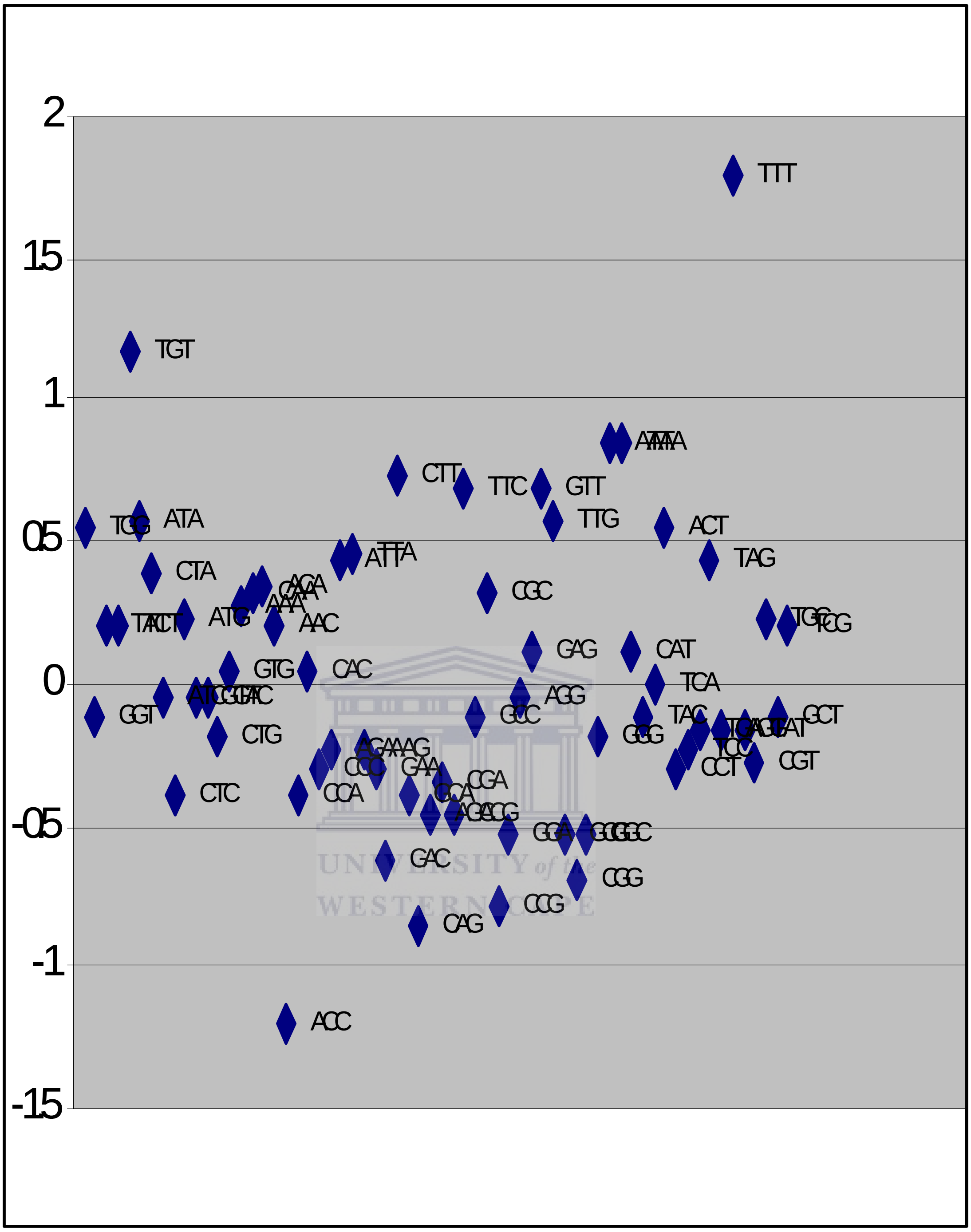
```
TCT =  0.1939 TGT = 1.16360 ATA =  0.5818
CTA =  0.3879 ATC = -0.0388 CTC = -0.3879
ATG =  0.2327 GTA = -0.0388 GTC = -0.0388
CTG = -0.1939 GTG =  0.0388 AAA =  0.2715
CAA =  0.3103 ACA =  0.3491 AAC =  0.1939
ACC = -1.2024 CCA = -0.3879 CAC =  0.0388
CCC = -0.3103 AGA = -0.2327 ATT =  0.4267
TTA =  0.4655 AAG = -0.2327 GAA = -0.3103
GAC = -0.6206 CTT =  0.7370 GCA = -0.3879
CAG = -0.8533 AGC = -0.4654 CGA = -0.3491
ACG = -0.4654 TTC =  0.6982 GCC = -0.1164
CGC =  0.3103 CCG = -0.7758 GGA = -0.5430
AGG = -0.0388 GAG =  0.1164 GTT =  0.6982
TTG =  0.5818 GCG = -0.5430 CGG = -0.6982
GGC = -0.5430 GGG = -0.1939 AAT =  0.8533
TAA =  0.8533 CAT =  0.1164 TAC = -0.1164
TCA =  0.0010 ACT =  0.5430 CCT = -0.3103
TCC = -0.2327 TGA = -0.1551 TAG =  0.4267
AGT = -0.1552 TTT =  1.7842 GAT = -0.1552
CGT = -0.2715 TGC =  0.2327 GCT = -0.1164
TCG =  0.1939
```

UNIVERSITY of the
WESTERN CAPE

Figure 2.3.2. A hash table of scores/figures generated from a promoter/non-promoter dataset pair. Each dataset (promoter or non-promoter) consists of 50 sequences of 55 bp sequence-length each. The actual frequency value of each triplet in the promoter set is subtracted from its corresponding value in the non-promoter (equation 2.3.2) to generate the hash table values. Certain triplets in the hast table have relatively high values. For example, TAA, TGT and TTT, an indication that, they are more prevalent in the set of promoter sequences as compared to non-promoter sequences (coding sequences). Similarly, other triplets

with negative scores are generally more prevalent in the non-promoter (coding sequences) as compared to the promoter sequences e.g. CAG.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 2.3.3. A scatter plot of hash table of scores/figures generated from a promoter non-promoter pair shown in figure 2.3.2.

Cumulative score and therefore the performance of a test sequence is assessed by :

a. Opening a 3-bp window and extracting all the triplets in the sequence as the window is shifted 1 bp to the end.
b. Obtaining each triplet's corresponding hash table value.
c. Summing up the scores of all the hash table values that corresponds to the triplets found in the sequence.

For a given set of sequence *S*, the frequency f of each triplet is determined by:

$$f_{\text{triplet}} = \frac{(Ns_t)(4^3)}{M_s} \qquad 2.3.1.$$

Where Ns_t represents the number of times a particular triplet occurs in the sequence set *S*, M_s is the total number of nucleotides in the set *S*.

UNIVERSITY *of the*
WESTERN CAPE

Hash table values for each triplet are obtained by:

$$\Delta f_a = f_a^P - f_a^{NP} \qquad 2.3.2$$

Where *P* and *NP* represent promoter and non-promoter respectively and Δf_a *represents* the hash table value of a particular triplet α.

2.2.3. Scoring on test sequences.

Since hash table values are obtained by subtracting the frequency of a triplet found in non-promoters from that of the corresponding value in promoters, the higher the value, the greater the likelihood of the sequence in question being a promoter. Triplets found to be present in almost equal numbers in both promoters and non-promoters almost cancel out and therefore have practically no contribution to the score(s). Each test sequence is assessed, by adding up the hash values of the triplets as a 3-bp window is shifted 1 bp until the end of the sequence. It must be noted that, the score itself is meaningless unless it is compared to a cut-off or a threshold value. Such a cut-off score would have to be obtained from a group of known promoters tested on the same hash table. Examples are found in chapter five, where TFDA has been used to analyze and predict promoters and non-promoters of *E.coli*, *B.subtilis* and *M.tuberculosis*.

UNIVERSITY of the
WESTERN CAPE

Chapter three.

Using Hidden Markov Models/Profiles on *E.coli*, *B.subtilis* and *Mycobacteria* promoters.

ABSTRACT

Hidden Markov Models for 'promoter sequence family' were implemented on sequence data from *E.coli, B.subtilis* and Mycobacteria. These implementations are based on the assumption that: features of promoter regions that are determinants for directing RNA polymerase to the binding site must be present as conserved elements. Promoter profile-like HMMs were therefore built/developed on various subsets of promoter sequences from *E.coli, B.subtilis* and Mycobacteria. The different promoter models (profiles) were then tested on separate datasets of promoter and non-promoter sequences to determine how the individual profiles/models discriminated promoter against non-promoter sequences. Results from the study revealed that, HMM models trained/built on promoters were capable of predicting/detecting other promoter sequences (not exposed to training) from non-promoter (coding) sequences effectively. Encouraging results of 90% true positive (TP) to a low false positive (FP) prediction of ~6% and ~3% were achieved for *E.coli* and *B.subtilis* data respectively. The results (~13% FP) obtained from similar studies on Mycobacteria promoters were not as encouraging as those obtained on *E.coli* and *B.subtilis*. Insufficient training data as well as 'dirty' training and test data set among others could have been contributory factors to the poor results on Mycobacteria test data.

3.1. Introduction

Computational analysis is increasingly becoming important for inferring functions and structures of regulatory sequences and proteins. Apart from large volumes of sequence data being generated, increase in computational power and readily available information over the internet are some of the reasons why biological science is gearing towards the direction of biocomputation. In this chapter, computational analysis using Hidden Markov Model (HMM) is applied in detection and prediction of prokaryotic promoters. Proteins, RNAs and regulatory sequences can usually be classified into families of related sequences and structures (Henikoff *et al.*, 1997). Ordinarily, sequence alignment would reveal functional relatedness between the families of related sequences such as promoters. However, the complex nature of promoters coupled with their variety and size(s) necessitates the inclusion of position-specific information from multiple alignments when searching for similar sequences. Pairwise sequence comparison algorithms such as BLAST and FASTA were designed based on the assumption that, all positions are equally important. However, great deal of position-specific information is usually available to the sequence families. Profile methods for building position-specific scoring models from multiple alignments were introduced for such purposes (Taylor, 1986; Gribskov *et al*., 1987; Barton, 1990; Henikoff, 1996). A ‘profile’ is defined as a consensus primary structure model consisting of position-specific residue scores and insertion/deletion penalties. Hidden Markov Models (HMMs) provide a coherent theory for profile methods (Henikoff, 1996). Profile HMMs have already been employed in many biological applications including protein modeling (Krogh *et al*., 1994; Baldi and Chauvin, 1995), gene prediction (Borodovsky *et al*., 1995;

UNIVERSITY of the
WESTERN CAPE

Lukashin and Borodovsky, 1997) and promoter studies (Yada *et al*., 1996; Pedersen *et al*., 1996; Lazareva-Ulitsky *et al*., 1999).

Regardless of the training method, once HMM has been successfully trained on a family of sequences, it can be used in a number of different tasks. First, for any sequence, one can compute the likelihood of the sequence in question to the fit the model. The trained model can also be used in discriminatory test and database searches (Krogh *et al*., 1994; Baldi and Chauvin, 1994) by comparing the likelihood of any sequence to model the sequences in the family on which an HMM model has been developed. Finally, the parameters of a model, such as emission distributions of the backbone (main) states and their entropies can be used to detect consensus patterns and other signals (Baldi *et al*., 1995). In this study, various hidden Markov profiles are modeled on different sequence sets (varied number of sequences of various fragment sizes) to study how well HMM can model on various sequence numbers and sizes. The developed models are then tested on their ability to discriminate against non-similar sequences. HMM is selected for this study because of the properties mentioned above and also due to its availability in the form of HMMer (Eddy, 1997).

UNIVERSITY of the
WESTERN CAPE

3.2. METHODS

3.2.1.1. *E.coli* Promoter Sequences.

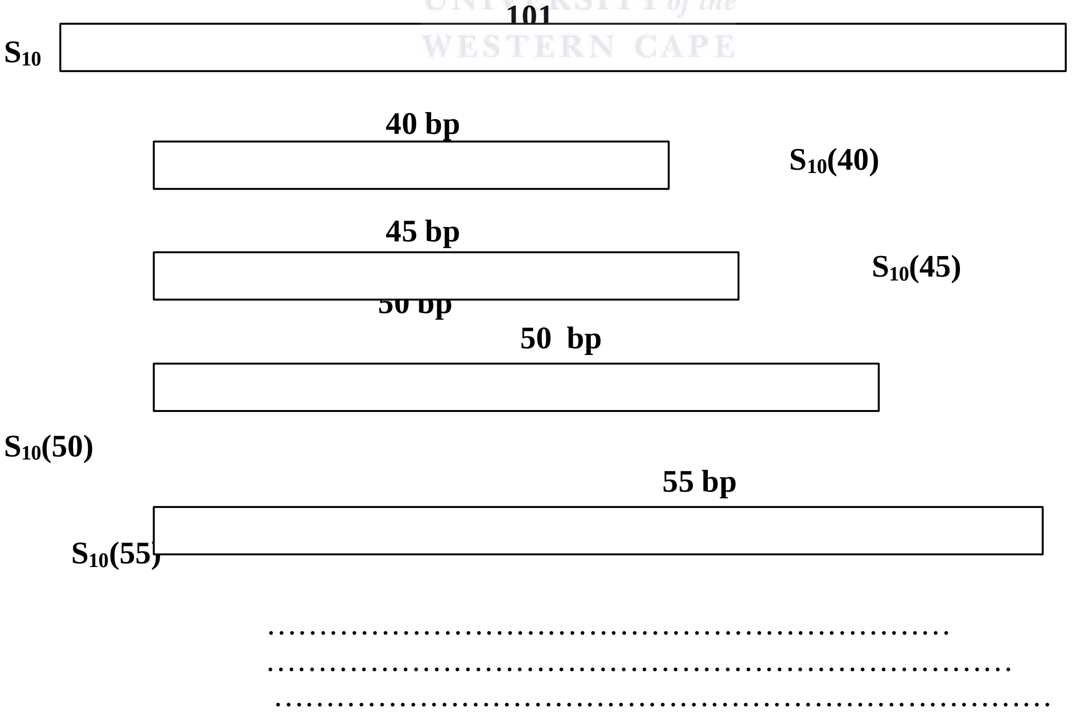
E.coli promoter sequences were taken from the dataset compiled by Lisser and Margalit, (1993). The total number of promoter sequences in this dataset is 300. Most of the promoter sequences in the database have sequence length of 101 bp (75 bp to tss and 26 bp after tss). However, there were a small number promoter sequences with smaller number of nucleotides in the promoter dataset t. Annotated *E.coli* genome sequences obtained from Genbank (version 111) were used to extend shorter promoter sequences to 101 bp using the respective tss as reference site. For example, if a promoter sequence consisted of 73 bp up to the tss, two nucleotides were added to the 5’ end of the sequence and 26 nucleotides to the 3’ of the sequence. Overlapping promoter sequences and promoter sequences with multiple or unconfirmed transcriptional start sites were removed from the data. The resulting set consisted of 168 promoters. The 168 promoters were randomly divided into two sets with no regard to relationships between specific promoters and their sigma (σ) factors. The first set of 83 promoters (Appendix_one) was used for training whilst the other set (Appendix_two) was used to test the performances of the various algorithms.

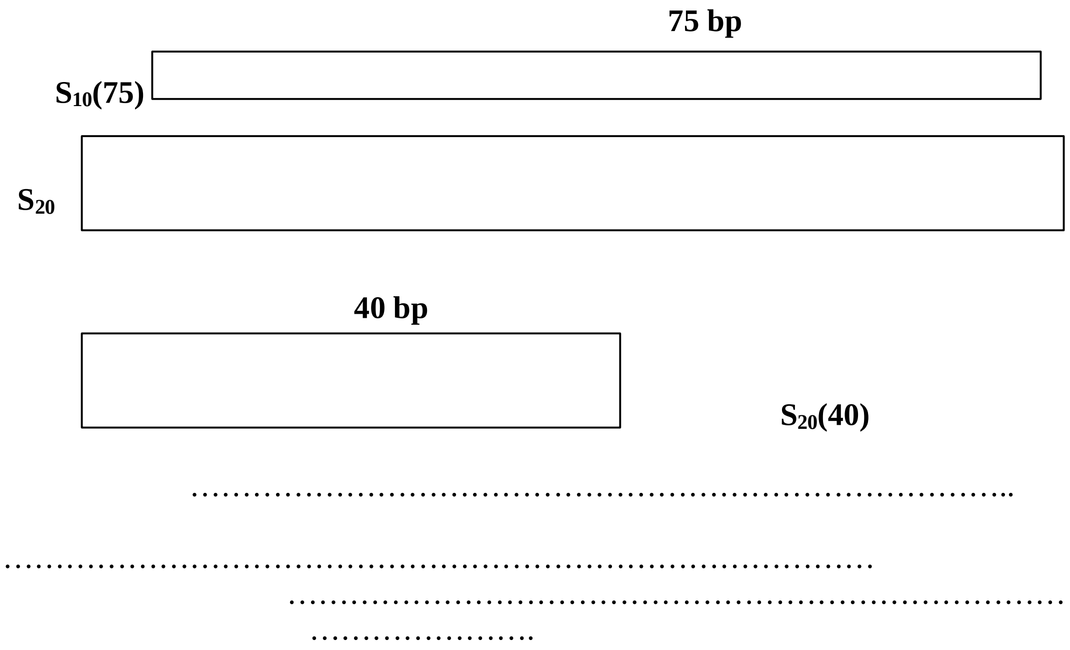
3.2.1.1.1. Generation of sequence sets for modeling.

UNIVERSITY of the
WESTERN CAPE

To generate promoter sequence subsets for HMM modeling, sequence sets comprising of ten (S_{10}) to fifty (S_{50}) sequences were randomly generated from the 83 promoter sequences making up the training set. Each sequence subset (S_{10}- S_{50}) subset was further sub-grouped according to sequence length that ranged from 40 bp to 75 bp (figure 3a). In all cases, promoter sequence subsets with sequence length up to 50 bp consisted of the transcription start site and the immediate upstream sequences, that is, -50 to tss. Promoter sequence subsets with sequence lengths greater than 50 bp used in training had the first 50 bp selected from upstream of tss (inclusive) with the only exception being on sequences of 75 bp sequence length (-55 to +20).

UNIVERSITY of the
WESTERN CAPE





UNIVERSITY of the
WESTERN CAPE

Figure 3a. A diagram depicting how various sequence subsets were generated from the original training dataset of 83 promoters. The diagrams representing sequence sets are not drawn to scale.

3.2.1.2. *E.coli* Non-Promoter Data

E.coli non-promoter sequences were generated from *E.coli* coding sequence file 'ecoli.ffn' obtained from Genbank (version 111). Sequence lengths of 101 bp were extracted from randomly selected coding sequences in the Genbank file 'ecoli.ffn'. Datasets similar to those of the promoter sequences were generated. Five thousand (5000) of the selected coding sequences (Appendix_three) were used as test sequences. All selected coding sequences were manually screened to ensure that, they did not contain any known *E.coli* promoter sequences. However, there is no disputing that they could probably contain promoter(s) not yet characterized though its quite unlikely considering the number (83).

3.2.2.1. *B.subtilis* Promoter Data.

B.subtilis promoter sequences were obtained from two sources. Promoters transcribed by sigma factor *A* (*sA*) were obtained from a compilation by Helmann (1995). Promoters transcribed by other sigma factors (σB, σC, σD, σE, σF, σG, σH, σK and σL) were obtained from the compilation by Yada *et al*., (1997). Promoter sequences with experimentally unconfirmed tss and multiple tss were removed from the dataset. Annotated *B.subtilis* genomic data (Genbank release 111) were used to extend each of the selected sequences to 101 bp each, 75 bp upstream of tss (inclusive) and 26 bp downstream of tss. The selected promoter sequences (164), were randomly divided into two sets of 83 and 81 sequences. The set of 81 was used in training/building all the different HMM profiles (Appendix-_four). Promoter sequence subsets were generated in a similar manner to *E.coli* (section 3.2.1.1.1). The other set of eighty-one (83) promoters (Appendix_five) was used as test data.

UNIVERSITY of the
WESTERN CAPE

Type of sigma factor	**Symbol**	**Number of promoters used**
SigmaA	**σA**	81
SigmaB	**σB**	8
SigmaC	**σC**	4
SigmaD	**σD**	7
SigmaE	**σE**	25
SigmaFG	**σ F** and **σG**	15
SigmaH	**σH**	9
SigmaK	**σK**	9
SigmaL	**σL**	3

Table 3.1. The source of the 162 *B.subtilis* promoter sequences that were split into two sets (training and testing promoter data). Promoter sequences were obtained from Helmann (1996) and Yada *et al.*, (1997). Sequences were thoroughly shuffled (no compromise on which promoters are transcribed by which sigma factors) before being divided into the two sets i.e. training and test data.

UNIVERSITY of the
WESTERN CAPE

3.2.2.2. *B.subtilis* Non-Promoter Data

B.subtilis non-promoter sequences were generated from *B.subtilis* coding sequences ‘bsub.ffn’; (Genbank version 111). Sequence lengths of 101 bp were extracted from randomly selected coding sequences in the Genbank file ‘bsub.ffn’. Datasets similar to those made for *E.coli* non-promoter sequences were created (Appendix_six). The data sets were used in testing the ability of the built/developed promoter profiles to discriminate against non-promoter sequences. Selected non-promoter data were screened for known *B.subtilis* promoter sequences as with *E.coli* non-promoter data.

3.2.3.1. *M.tuberculosis* promoters

M.tuberculosis promoter sequences were obtained from several publications (Appendix_ seven). Only promoters with experimentally characterized transcriptional start sites (tss) were used in the training set. Altogether, 26 *M.tuberculosi*s promoter sequences were obtained with established transcriptional start site (tss). Twenty-four (24) other mycobacterial promoters (Appendix_eight) were added to the initial 26 to constitute the training set. Where possible, mycobacterial genome data was used to extend the promoter sequences to 101 bp (75 bp to tss and 26 bp after tss). Thirty-three (33) other mycobacterial promoters (Appendix_nine) with unknown tss but known –10 or –35 were selected as the test data. *M.tuberculosis* genome data was used to fill in such sequences to101 bp depending on which of the two canonical hexamers (–10 or –35) was known. For example, a promoter sequence up to –10 hexamer was extended by about 33 bp(+7 to tss and +26 after tss) and the 5’ end adjusted to make up the 101 bp.

UNIVERSITY of the
WESTERN CAPE

3.2.3.2. *M.tuberculosis* Non-promoter Data

M.tuberculosis coding sequences were used as non-promoter data. Coding sequences from the Genbank file `mtub.ffn’ were randomly selected and sequence lengths of 101 generated from them. Data sets similar to those made for *E.coli* and *B.subtilis* promoters were created and used to determine the discrimination ability of the individual models/profiles. Total number of coding sequences used for testing was five thousand (5000). The *M.tuberculosis* non-promoter dataset can be found in Appendix_ten. Non-promoter data was screened for any known mycobacterium promoters in the same manner as those of *E.coli* and *B.subtilis* non-promoter data.

3.2.4. HMMER software

The HMM software used is this research is the HMMer package version 1.8 developed by Sean Eddy (Eddy, 1995). HMM models were built for each subset from the promoter data (10-45 to 50-75) using *hmmt* (hmmtrain). `The program, '*hmmt'* learns patterns shared by multiple sequences and saves the pattern in *hmmfile*. *Hmmt* works by iteratively improving a new sequence alignment calculated using the model, then a new model using the current alignment. To avoid or minimize bad local minima in the training process, simulated annealing is used in the optimization of the alignments. *'Hmma'* (hmmaligh with a score option) which produces scores based on how well the sequence fits/aligns to the built model/profile, was used to categorize test sequences. Each specific model was used to test promoter and non-promoter sequences that corresponded to the model with respect to sequence length. Other tests were carried out on 75 bp fragment sizes and the entire sequence length of 101 bp. The latter tests were done by opening a sequence window that had same size as the particular model being used and obtaining the cumulative score as the window is shifted one bp (figure 3.l.5).

UNIVERSITY of the
WESTERN CAPE

The HMMer package was compiled on a SGI irix workstation (irix 6.3).

3.2.5. Scoring with HMM.

Respective HMM models trained on the various promoter sequence subsets, (section 3.2.1.1.1) were first tested on test promoter sequences having the same sequence length as the sequences used to develop the models. Each sequence produces a score when tested on the corresponding model. The higher the score, the more the sequence 'fits' the HMM model that was used to test the sequence. The promoter test sequence scores from each category (fragment/sequence length) length were then arranged in descending order with respect to the value of the scores. Scores from each promoter test data that resulted in 90% true positives (TP) were used as threshold values to categorize test sequences (coding sequences) as predicted promoters/non-promoters.

In the other test cases, where test sequences (coding) were of fixed lengths (75 bp and 101 bp), the same procedure was applied to the promoter test data to obtain threshold values that resulted in 90% true positive (TP). Depending on which HMM model used, window sizes equivalent to the size used to develop the models were opened in the test sequence(s) and the window(s) shifted a bp until the end of the test sequence as illustrated in figure 3.1. Predicted results were summed up and arranged in descending order starting from the highest value as above to obtain the threshold values.

WESTERN CAPE

3.3. Results and Discussion

Hidden Markov Models (HMM) profiles/models of promoter sequences were successfully developed by training them on different sets of promoters from *E.coli*, *B.subtilis* and Mycobacteria. The promoter sequences used in training the models were aligned according to their respective transcriptional start sites (tss). Promoter training datasets ranged from ten (10) sequences of 40 bp sequence length (S_{10}(40) or 10_40) to fifty sequences of 75 bp sequence length 50_75 (S_{50}(75) (as in section 3.2.1.1.1). Eighty-three (83) separate promoter sequences used to test the performance of the models in both *E.coli* and *B.subtilis*. However, only thirty-four (34) of such sequences were available for the study on Mycobacteria. In all the three cases, five thousand (5000) sequences extracted from their respective coding sequences were used as non-promoter test data. Since a major feature of promoters (both eukaryotes and prokaryotes) is what appears to be multiple signal covering the entire promoter region, three types of tests as described in the protocol section were carried out with the individual promoter models developed from HMM. In the first designed test, (test *A*), test sequences had the same sequence length as the corresponding data used to train/develop models. Test *B* was performed on sequences of 75 bp fragment sizes (promoters and non-promoters), whilst test *C* was performed on 101 bp sequences (see fig. 3.1). The composition of all the promoter datasets used in training and testing is as follows: Nucleotide sequences with fragments up to 50 bp were selected upstream of the transcription start sites inclusive. Promoter sequences with fragment sizes greater than 50 bp had the extra nucleotides selected after the transcriptional start site. For example, a promoter sequence fragment of 65 bp would consist of 50 bp nucleotides upstream to the transcriptional start site and 15 bp downstream to the transcriptional start site (-50 to +15).

UNIVERSITY of the
WESTERN CAPE

3.3.1. *E.coli*

Not all of the promoter sequence sets were successfully `profiled' on HMM. Those unsuccessful sets include 20_75 in the sets of twenty sequences, 30_75 in the set of thirty sequences, 40_65, 40_70, 50_75 in the set of forty and 50_65, 50_70, and 50_75 for the set of fifty sequences. Normally, hmmt generates models after thirty to sixty iterations depending on the number and fragment size of sequences. Inability to develop a model is tantamount to not being able to reach some kind of consensus on the sequence sets, which usually happens when sequence set is large with respect to number and fragment size. Most of the sequence sets that were not 'profiled' have sequence length from 70 bp upwards. Since the first 50 bp are selected upstream of transcriptional start site inclusive, it is logical to assume that, the extra sequences (+ 20 bp) after tss are responsible for the difficulty in obtaining profiles on the sequence sets. Successful model/profiles were tested on known *E.coli* promoters (83) and coding sequences (5000) to determine which profile best represented the information harbored in the promoter training set. In all test cases using the eighty-three (83) promoter test sequences, individual cut-off scores that resulted in ~90% (75/83) true positive were used to categorize the test sequences (promoters or non-promoters). Tables 3.1, 3.2 and 3.3 show the results of various promoter-trained models on five thousand (5000) non-promoter sequences of same length as models, 75 bp and 101 fragment sizes respectively. Because of the problem of which position to start from when selecting different fragment sizes from the original 101 bp non-promoter, five (5) test sequences were generated from each original test sequence for fragment sizes less than 75 bp inclusive. The averages from these five results were adopted as the results for each test sequence. Plots of the results of the individual false positives obtained from the different HMM promoter models on coding sequence (CDS) of same size as model, 75 bp sequences and 101 bp sequences are shown in figures 3.2, 3.3 and 3.4 respectively.

```
A.

GATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCG
GATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAA                              -13.556
.......
 ATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAA                             -19.514          -
33.070
  TCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAG                            -21.733          -
54.803
   CACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGA                           -15.378          -
70.181
    ACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAG                          -15.260          -
85.441
     CACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGG                         -17.535         -
102.976
      ACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGT                        -21.812         -
124.788
       CAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTT                       -21.467         -
146.255
        AAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTG                      -16.482         -
162.737
```

UNIVERSITY of the
WESTERN CAPE

```
B.

GATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCGTTTAGGTGTTTTCACGAGC
ACTTCA
GATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAA                              -13.556
.......
 ATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAA                             -19.514          -
33.070
  TCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAG                            -21.733          -
54.803
   CACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGA                           -15.378          -
70.181
    ACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAG                          -15.260          -
85.441
     CACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGG                         -17.535         -
102.976
      ACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGT                        -21.812         -
124.788
       CAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTT                       -21.467         -
146.255
        AAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTG                      -16.482         -
162.737
```

Fig. 3.1. A diagrammatic illustration of how a trained model was used to test fragment sizes of 75 bp (A) and 101 bp (B). Individual results (column2) and cumulative results (column 3) obtained from a model trained on a set of thirty sequences with forty-five bp fragment size (30_45) on a test sequence of fragment length 75 bp (A) and 101 bp (B). A moving window of 45 bp is opened from the first nucleotide and shifted one bp till the end. The scores from alignment of each window to the trained model and the cumulative scores are shown on the second and third columns respectively. Cut-off scores that generated 90% true positive were selected to determine whether the sequences under investigation are adjudged as promoter(s) or not. The only difference between the two test sequences above is that additional scores are generated for the 101 bp test sequence.

UNIVERSITY of the WESTERN CAPE

Set	1	2	3	4	5	Average	%FP
10_40	1195	1171	1176	1131	1176	1170	23.4
10_45	846	846	854	794	846	837	16.7
10_50	832	850	832	866	864	849	17.0
10_55	1273	1277	1269	1239	1260	1263	25.3
10_60	922	926	920	859	938	913	18.3
10_65	1142	1123	1099	1122	1168	1131	22.6
10_70	1072	1070	1068	1049	1076	1067	21.3
10_75	1211	1251	1253	1235	1237	1237	24.7
20_40	842	861	859	806	879	849	17.0
20_45	1003	1034	965	1003	1020	1005	20.1
20_50	728	721	740	749	719	731	14.6
20_55	1551	1599	1541	1588	1571	1570	31.4
20_60	612	612	607	596	592	604	12.1
20_65	1426	1431	1410	1441	1393	1420	28.4
20_70	997	981	1030	957	1025	998	20.0
20_75	-	-	-	-	-	-	-
30_40	699	690	667	663	694	683	13.7
30_45	642	664	643	636	681	653	13.1
30_50	725	689	696	721	712	709	14.2
30_55	1364	1367	1370	1370	1404	1375	27.5
30_60	625	643	663	645	639	643	12.9
30_65	1113	1138	1107	1118	1102	1116	22.3
30_70	668	671	699	667	678	677	13.5
30_75	-	-	-	-	-	-	-
40_40	459	466	442	445	446	452	9.0
40_45	385	378	145	356	398	332	6.6
40_50	454	464	479	479	450	465	9.3
40-55	767	797	783	794	818	792	15.8
40-60	843	840	817	856	798	830	16.6

40-65	-	-	-	-	-	-	-
40-70	-	-	-	-	-	-	-
40_75	-	-	-	-	-	-	-
50_40	591	646	580	547	601	593	11.9
50_45	578	598	578	601	627	596	12.0
50_50	830	805	781	799	820	807	16.1
50_55	613	640	638	648	654	639	12.8
50_60	685	690	685	726	681	693	13.9
50_65	-	-	-	-	-	-	-
50_70	-	-	-	-	-	-	-
50_75	-	-	-	-	-	-	-

Table 3.2. Number of false positives obtained for HMM trained models on promoter subsets. Sequences used in testing both promoters and non-promoters had the same number of nucleotides as those used in development of the models. Those sequence sets which could not be trained using HMM are marked with ‘-‘. Five sub-fragments were generated from each test sequence. Depending on the sequence length of sub fragments, the first nucleotide is chosen randomly within the possible range that would make the fragment size possible in the 101 bp sequence. The averages from the five sub-fragments and the corresponding percentage false positives are shown on the sixth and the seventh columns respectively. All promoter and non-promoter data are from *E.coli*.

UNIVERSITY of the
WESTERN CAPE

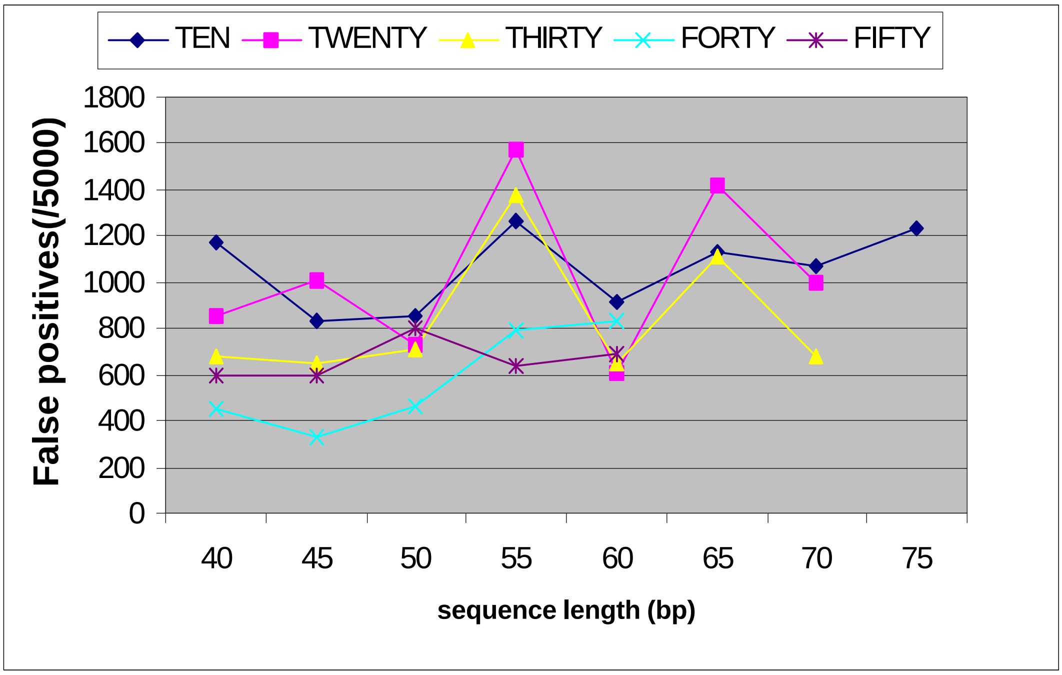


Fig. 3.2. Individual trained HMM models with their corresponding false positive results on 5000 coding sequences. Model 40_45 (forty promoter sequences of 45 bp sequence each) produced the best results (least number of false positives - 385). Models were tested on sequences having same fragment sizes as those used in building the models. A cut-off score that produced 90 % (75/83) True positive (TP) was used to select the predicted promoters from non-predicted promoters. Thus in all cases, true positive rate is ~90%.

Sets	**1**	**2**	**3**	**4**	**5**	**Av.**	**%**
10_40	753	760	751	742	764	754.0	15.1
10_45	710	719	695	696	701	704.2	14.1
10_50	719	710	711	724	715	715.8	14.3
10_55	771	740	724	758	786	755.8	15.1
10_60	898	900	937	914	909	911.6	18.2
10_65	1282	1306	1301	1291	1313	1298.6	26.0
10_70	1289	1267	1284	1281	1271	1278.4	25.6
10_75	1551	1573	1545	1553	1537	1551.8	31.0
20_40	609	612	636	632	608	619.4	12.4
20_45	687	682	692	701	676	687.6	13.8
20_50	457	471	472	454	484	467.6	9.4
20_55	759	758	738	763	757	755.0	15.1
20_60	694	677	700	709	694	694.8	13.9
20_65	1236	1257	1231	1243	1274	1248.2	25.0
20_70	910	939	899	923	892	912.6	18.3
30_40	623	611	597	605	599	607.0	12.1
30_45	493	492	498	495	483	492.2	9.8
30_50	513	484	512	505	489	500.6	10.0
30_55	710	687	694	697	693	696.2	13.9
30_60	800	793	781	799	775	789.6	15.8
30_65	501	523	514	516	513	513.4	10.3
30_70	736	731	726	741	721	731.0	14.6
30_75	-	-	-	-	-	-	-
40_40	527	540	541	531	518	531.4	10.6
40_45	386	395	405	381	378	389.0	7.8
40_50	527	508	530	520	529	522.8	10.5
40_55	495	509	485	507	495	498.2	10.0
40_60	557	542	553	566	558	555.2	11.1
40_65	-	-	-	-	-	-	-
40_70	-	-	-	-	-	-	-
40_75	-	-	-	-	-	-	-

UNIVERSITY of the WESTERN CAPE

50_40	445	468	451	452	444	452.0	9.0
50_45	360	362	374	374	357	365.4	7.3
50_50	370	380	372	387	362	374.2	7.5
50_55	446	464	451	462	442	453.0	9.1
50_60	487	469	483	488	474	480.2	9.6
50_65	-	-	-	-	-	-	-
50_70	-	-	-	-	-	-	-
50_75	-	-	-	-	-	-	

Table 3.3. Number of false positives obtained for HMM trained models on promoter subsets. Nucleotide sequences used for testing both promoters and non-promoters had the constant sequence length of 75 bp. Five different sequences were generated from each test sequence of 101 bp. The first nucleotide of each of the five sets was selected randomly from nucleotide number one (1) to twenty-six (26). Individual performances (non-promoters) were obtained by moving a window within the 75 bp that corresponds with the model and summing up the scores as the window is shifted one bp, fig 3.1. Sequence sets that could not generate HMM profiles are marked with '-'. The average and the percentage false positives are shown on the sixth and the seventh columns respectively. The above results were obtained on 90% true positives.

UNIVERSITY of the
WESTERN CAPE

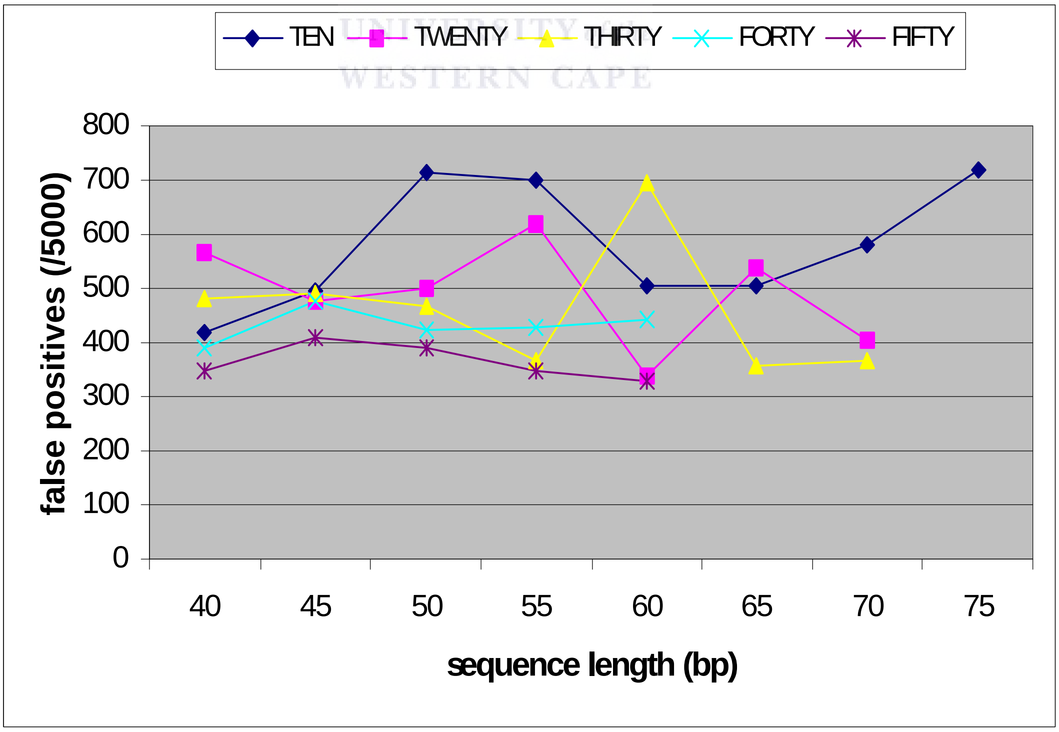


Fig. 3.3. Individual HMM sequence models with corresponding false positive results on 5000 coding sequences of 75 bp sequence-length each. Each sequence's score was obtained by opening a window within the 75 bp sequence, which corresponded to the model size, and summing the results as the window was shifted 1 bp, fig. 3.1. As in the previous case, scores that resulted in 90% true positive from the 83 promoters were used as the cut-off score to distinguish between predicted promoters and non-promoters.

UNIVERSITY of the
WESTERN CAPE

A

	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	421	566	480	391	346
45	496	474	492	474	409
50	714	498	468	422	390
55	699	621	365	427	350
60	504	338	694	444	330
65	506	536	355	-	-
70	583	404	365	-	-
75	72	-	-	-	-

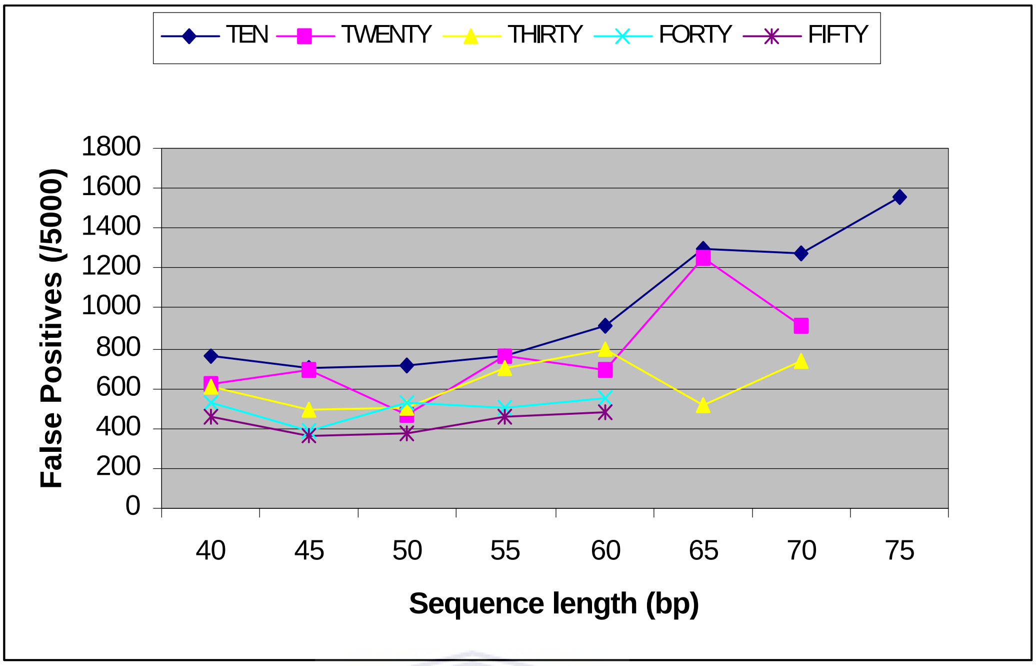
0

B

	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	8.4	11.3	9.6	7.8	6.9
45	10.0	9.5	9.8	9.5	8.2
50	14.3	10.0	9.4	8.4	7.8
55	14.0	12.4	7.3	8.5	7.0
60	10.1	6.8	13.9	8.9	6.6
65	10.1	10.7	7.1	-	-
70	11.7	8.1	7.3	-	-
75	14.4	-	-	-	-

Table 3.4A. Number of false positives obtained from the HMM models trained on the different subsets of *E.coli* promoter sequences. Promoter and non-promoter (coding sequences) fragment sizes of 101 (fig. 3.1.B) were used in the test. Threshold values that resulted in 90% true positives (TP) were used. Rows marked '-' indicate promoter subsets that could not be trained or modeled successfully on HMM. Test sequence values were obtained as in figure 3.1. Table 3.4B is in percentages instead of actual numbers.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Fig. 3.4. Individual HMM sequence models with corresponding false positive results on 5000 coding sequences of 101 bp sequence-length each (test sequence). Each sequence's score was obtained by opening a window within the 101 bp sequence, which corresponded to the model size, and summing the score as the window was shifted 1 bp, fig. 3.2. Threshold scores that resulted in 90% true positives from the 83 promoters were used.

Training of the individual promoter sets was performed iteratively until the best models/profiles were achieved with respect to how well the promoter profiles/models were able discriminate against non-promoter test sequences. The results from the three tests, figures 3.2, 3.3 and 3.4 suggest that, false positive rates obtained from the sequence sets get better (less false positives), with increase in the number of promoter sequences used for training. Good results were obtained from models trained on sets of forty and fifty sequences. With greater number of sequences available for training, the model will more reliably capture statistical properties of the training set. No correlation is apparent between score and fragment size for any particular set of sequence. Results from the study on *E.coli* suggest that, trained HMM models do not necessarily improve (as measured by the ability of the models to discriminate promoter against non-promoter sequences) with increase in sequence size of the promoter set. Of course, that would depend on which section of nucleotide sequence is taken to define promoter. In this study, the entire region of about 101 (76 bp up to tss to 25 bp after tss) has been under investigation as promoter region. Models appear to peak in performance (least number of false positives) around the region of +1 to -45 and +1 to -50 bp. The region about 20 bp from the transcription start site, between 50th nucleotide and 70th produced variable results for all sequence sets except for the sequence set made up of twenty sequences produced results/scores that are more variable scores. The best result of the study (least number of false positives) is observed on the model trained on a set of forty sequences with fragment sizes of 45 bp (40_45). The false positive (FP) score of 332 (6.7%) is relatively low compared with the next best score 465, also coming from the model/profile trained on forty promoter sequences of 40 bp sequence fragments selected upstream of their respective transcriptional start sites.

WESTERN CAPE

Unlike the scores obtained from testing sequences of the same sequence length as those used to build/train the models, the best results from both 75 bp and 101 bp sequence fragments were from models trained on sets of fifty sequences. Model trained on 50_45 (S_{50}(45)) produced the least number of false positives whilst 50_60 (S_{50}(60)) resulted in the best for all 101 bp test sequences. Certain promoter subsets produced results comparable to S_{50}(45) and S_{50}(60). They include models on 30_45 (S_{30}(45)) and 20_50 (S_{20}(50)) for the test on 75 bp sequences and 30_55 for the 101 bp test sequences. The results obtained on 101 bp sequences produced fewer false positives than those on 75 bp sequences, which were also better (less FP) than sequences of same size as models. These results provide support for the hypothesis that, promoter regions have multiple signals that are interspersed in the region upstream of -35 hexamer (Newlands *et al*., 1992). The best results obtained in each category, 6.7% FP (same test sequence length as models), 7.3% FP (fixed 75 bp test sequence lengths) and 6.7% FP (101 bp test sequences) are comparable to results obtained by researchers (Lukashin *et al*., 1989 (2-6%); O'Neill, 1992 (3-10%); Mahadevan and Ghosh, 1994 (8-10%) using other prediction methods. The true positive values of these researchers were also around 90%. The threshold value for all the test sequences was manually selected to give a true positive (TP) value of 90% for all promoter test sequences.

UNIVERSITY of the
WESTERN CAPE

3.3.2. *B.subtilis*

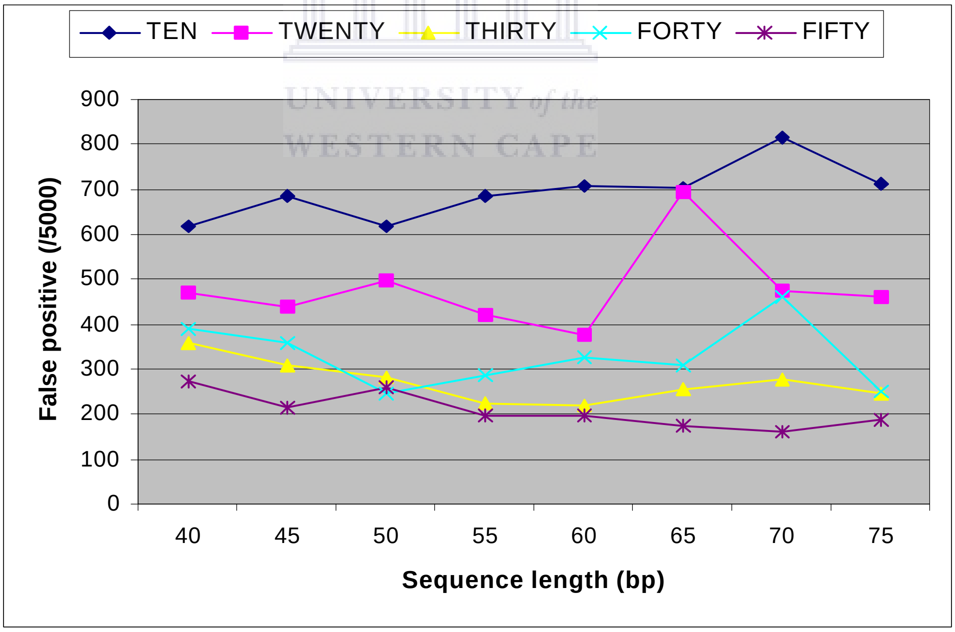
Having used *E.coli* promoter sequences on HMM to perform promoter predictions with some degree of success, the next task was to apply HMM modeling and prediction to another organism of significantly different nucleotide composition. This was to gain an insight into the degree to which nucleotide content or sequence variation would affect application of HMM in promoter predictability. In simple terms, would the results obtained on *E.coli* be different if for instance, the organism had higher or lower percentage GC composition in the organism's genome? Another significant challenge was to determine the minimal number of promoter sequences that could be successfully used on HMM model training; considering that, initial study had revealed that, fifty sequences produced very good results for *E.coli*. Consistency in results obtained from specific data sets would suggest that, the size in question would do well for other prokaryotes with different genomic composition with respect to the percentage GC content. *B.subtilis* was selected because it is a gram-positive bacterium differing distinctly from *E.coli* and *M.tuberculosis*, therefore a chance to study the concept on a different type of prokaryote, and also entertaining universality of the concept. The second reason is the easy availability of experimentally characterized *B.subtilis* promoters. The models were developed from an initial training set of 81 promoters and tested on 83 promoters. The experimental design was analogous to that used in the development of models for the *E.coli* HMM study. The true positive rate of every set was set to 90% by selecting scores that resulted in 90% true positives as threshold scores. The three types of tests as described earlier (test A, test B and test C) were also performed on the *B.subtilis* data. Five sequences were randomly generated from each non-promoter fragment for sequences less or equal to 75 bp, table 3.4 and 3.5. The averages from these five results were computed and adopted as the respective scores for the corresponding test sequences. The results of the individual false positives obtained from the different HMM models on *B.subtilis* coding sequences of same length as models, 75 bp fragment sizes and 101 bp are shown in figures 3.5, 3.6 and 3.7 respectively.

Sets	*1*	*2*	*3*	*4*	*5*	*Av*	*%*
10_40	689	642	580	596	580	617	12.3
10_45	743	659	672	702	658	687	13.7
10_50	685	579	619	591	628	620	12.4
10_55	762	642	676	671	669	684	13.7
10_60	790	689	663	681	714	707	14.1
10_65	763	670	700	695	678	701	14.0
10_70	888	804	819	795	781	817	16.8
10_75	764	715	706	676	696	711	14.2
20_40	499	478	457	448	472	471	9.4
20_45	493	415	421	431	443	441	8.8
20_50	555	479	490	478	493	499	10.0
20_55	471	399	402	402	430	421	8.4
20_60	465	354	373	349	348	376	7.6
20_65	808	675	685	663	634	693	13.8
20_70	562	444	441	460	459	473	9.4
20_75	524	458	433	440	454	462	9.2
30_40	396	343	347	363	349	360	7.2
30_45	359	306	283	279	317	309	6.2
30_50	369	264	272	262	254	284	5.6
30_55	254	209	206	209	232	222	4.4
30_60	295	207	217	179	205	221	4.4
30_65	290	246	231	256	244	253	5.1
30_70	346	277	268	246	251	278	5.6
30_75	310	221	231	234	237	247	4.9
40_40	415	394	395	386	358	390	7.8
40_45	419	332	354	328	364	359	7.2
40_50	271	225	244	240	245	245	4.9
40_55	344	253	280	282	283	288	5.8
40_60	399	303	298	329	306	327	6.5
40_65	351	309	287	313	295	311	6.2
40_70	550	440	436	433	452	462	9.2
40_75	296	233	246	245	241	252	5.0
50_40	301	261	268	260	265	271	5.4
50_45	234	218	221	200	209	216	4.3

UNIVERSITY of the
WESTERN CAPE

50_50	302	252	254	234	259	260	5.2
50_55	246	169	192	199	186	198	4.0
50_60	243	172	205	196	178	199	4.0
50_65	209	174	177	171	150	176	3.5
50_70	182	160	164	134	159	160	3.2
50_75	235	170	180	169	174	186	3.7

Table 3.5. Number of false positives obtained for HMM trained models on various promoter subsets. Nucleotide sequences used for testing both promoters and non-promoters had the same sequence length as sequence sets used in developing the respective models. Since there was a problem of which 75 bp windows of the 101 bp windows were to be used for testing, five different sequences were generated from each sequence with the nucleotide of the sequence being chosen randomly within the possible range in the 101bp with respect to the size of the sequence from which the models were built on. The average and the percentage false positives are shown on the sixth and the seventh columns respectively. Threshold scores were selected to have 90% true positive results for each test set.



UNIVERSITY of the
WESTERN CAPE

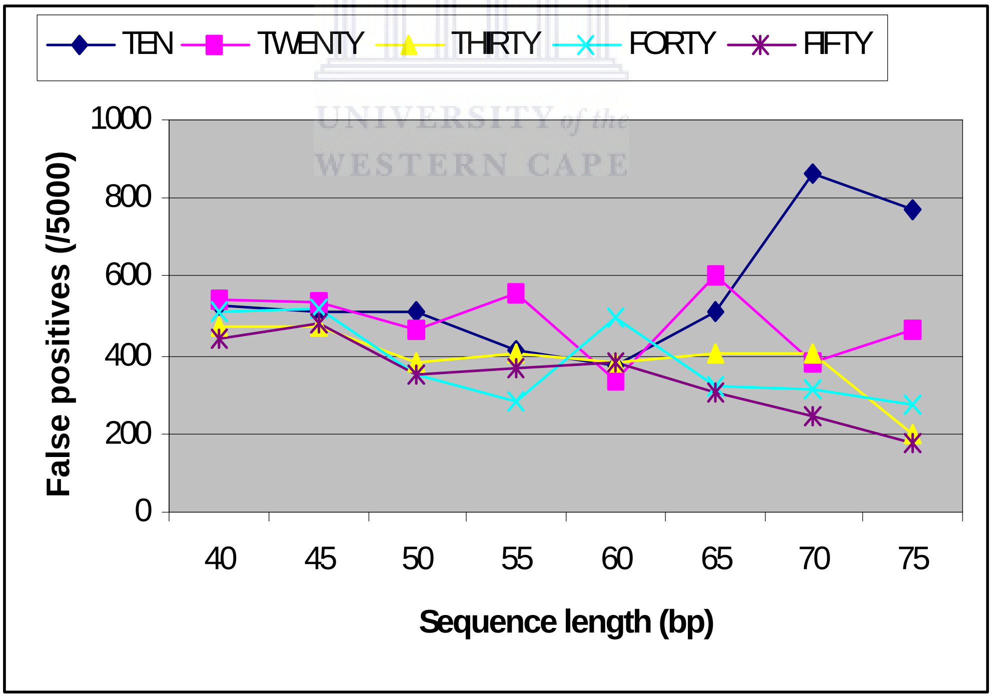
Fig. 3.5. Individual trained HMM models with their corresponding false positive results on 5000 *B.subtilis* coding sequences. Model 50_70 (fifty promoter sequences of fragment size 70 bp each) produced the best results (least number of false positives – 160). Models were tested on sequences having the same sequence length as those used in building the models. A cut-off score that produced 90 % (75/83) True positive (TP) was used to select the predicted promoters from non-predicted promoters.

Sets	*1*	*2*	*3*	*4*	*5*	*Av.*	*%*
10_40	636	504	495	509	482	525	10.5
10_45	616	504	483	494	480	515	10.4
10_50	598	491	485	477	489	508	10.2
10_55	495	396	378	381	393	409	8.2
10_60	455	361	338	360	339	371	7.4
10_65	602	504	480	494	479	512	10.2
10_70	945	839	854	825	833	859	17.2
10_75	825	769	776	723	751	769	15.4
20_40	662	510	505	527	498	540	10.8
20_45	653	515	494	500	495	531	10.6
20_50	578	438	428	438	427	462	9.3
20_55	694	522	529	522	514	556	11.1
20_60	405	318	325	319	324	338	6.8
20_65	711	594	577	576	570	606	12.1
20_70	474	354	371	361	367	385	7.7
20_75	531	457	433	443	453	463	9.3
30_40	588	463	440	434	430	471	9.4
30_45	592	455	446	438	420	470	9.4
30_50	475	363	367	360	340	381	7.6
30_55	485	365	378	390	386	401	8.0
30_60	468	352	361	361	348	378	7.6
30_65	508	382	378	387	382	407	8.2
30_70	501	383	386	362	371	401	8.0
30_75	243	180	172	183	204	196	3.9
40_40	629	483	474	481	479	509	10.2
40_45	650	499	485	484	485	521	10.4
40_50	428	347	331	334	317	351	7.0
40_55	354	270	273	264	265	285	5.7
40_60	622	468	455	470	480	499	10.0
40_65	392	303	289	295	312	318	6.4
40_70	388	298	312	294	286	316	6.3
40_75	322	254	267	261	271	275	5.5
50_40	543	428	414	424	423	446	9.0
50_45	596	463	456	451	428	479	9.6

UNIVERSITY of the
WESTERN CAPE

50_50	432	329	331	330	320	348	7.0
50_55	430	347	346	348	347	364	7.3
50_60	469	368	358	374	356	385	7.7
50_65	375	277	283	293	290	304	6.1
50_70	295	234	233	247	225	247	4.9
50_75	223	163	171	160	162	176	3.5

Table 3.6. Number of false positives obtained for HMM trained models on various *B.subtilis* promoter subsets. Nucleotide sequences used for testing both promoters and non-promoters had the same sequence length of 75 bp. Five different sequences were generated from each sequence with the nucleotide of the sequence being chosen randomly within the possible range in the 101bp with respect to the size of the sequence from which the models were built on. The scores were obtained by opening window within the 75 bp, which corresponds with the model, and summing up the scores as the window is shifted one bp, fig 3.1. The average and the percentage false positives are shown on the sixth and the seventh columns respectively.



UNIVERSITY of the
WESTERN CAPE

Fig.3.6. Individual HMM sequence models with corresponding false positive results on 5000 *B.subtilis* coding sequences of 75 bp sequence-length each. Each sequence's score was obtained by opening a window within the 75 bp sequence, which corresponded to the model size, and summing the results as the window was shifted 1 bp, fig. 3.1A. Scores that resulted in 90% true positive from the 83 promoters were used as the cut-off score to distinguish between predicted promoters and non-promoters.

UNIVERSITY of the
WESTERN CAPE

A

		SETS			
bp	TEN(%)	TWENTY	THIRTY	FORTY	FIFTY
40	591	599	518	638	507
45	646	743	644	728	640
50	674	673	551	503	654
55	575	667	513	526	498
60	666	768	535	721	579
65	630	607	566	567	547
70	844	623	732	509	671
75	727	624	534	566	622

B

		SETS			
bp	TEN(%)	TWENTY	THIRTY	FORTY	FIFTY
40	11.8	12.0	10.4	12.8	10.1
45	12.9	14.9	12.9	14.6	12.8
50	13.5	13.5	11.0	10.1	13.1
55	11.5	13.3	10.3	10.5	10.0
60	13.3	15.4	10.7	14.4	11.6
65	12.6	12.1	11.3	11.3	17.0
70	16.9	12.5	14.6	10.2	13.4
75	14.5	12.5	10.7	11.3	12.4

Table 3.7A. Number of false positives obtained on 5000 *B.subtilis* coding sequences of 101 bp sequence lengths. Threshold values that resulted on 90% *B.subtilis* promoter sequences were used. Fig. 3.6 shows the graph obtained from plotting the data. In table 3.7B, the false positive values are expressed as percentages. Figures in first column represent sequence length of the test sequences in the respective sets used for testing.

UNIVERSITY of the
WESTERN CAPE

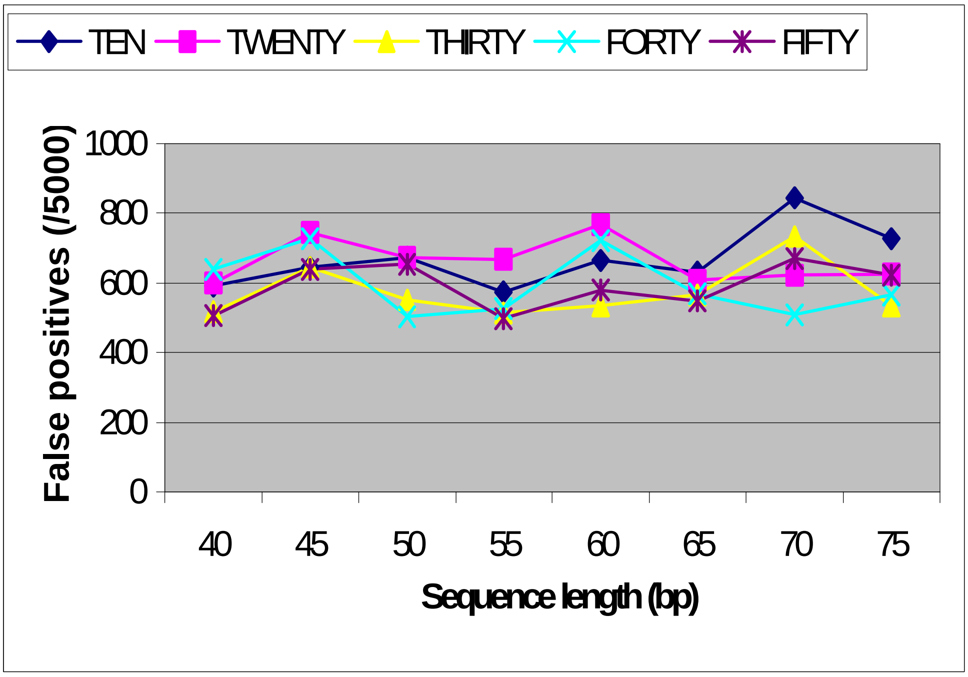


Fig. 3.7. Individual HMM models with corresponding false positive results on five thousand (5000) coding sequences of 101 bp fragment-size each. Each sequence's score was obtained by opening a window within the 101 bp sequence, which corresponded to the model size, and summing the score as the window was shifted 1 bp, fig. 3.1 B. Cut-off scores that resulted in 90% true positives from the 83 promoters were used.

UNIVERSITY of the
WESTERN CAPE

The results obtained from *B.subtilis* promoters and their corresponding non-promoters of coding sequences appear to be better than those obtained from *E.coli*. A best false positive rate of 3.2% for *B.subtilis* as compared to that of 6.7% for *E.coli*. There are some similarities between the two results. For example, promoters trained on smaller number of sequences (ten and twenty) produced inferior results as compared to promoters trained on more sequences (thirty, forty, and fifty). Also, most of trained models that seemed to discriminate best were on the sequence subsets with 55 fragment sizes (50 bp upstream of transcription start site inclusive plus 5 bp after tss). Obtaining a strong signal in the region 50 bp to the tss is once again expected as it harbors the conserved –10 and –35 hexamers. However, the extra five after tss is perhaps a region that needs to be carefully studied. In the same size as model/profile category, best results are obtained from sets of S_{50}(40) to S_{50}(75). However, the next best results are not from S_{40} set but rather S_{30} , in agreement with the earlier observation made on the study on *E.coli* promoters; increase in training does not necessarily always transform to better results. In the 75 bp test sequence category, the results (false positive) are more 'clustered' compared to the others. Best results are from this category (50_75). A close parallel relationship in scores with regard to false positives, seems to exist between set of S_{30} and S_{50} in both 75 and 101 bp test categories. The results seem to suggest that, HMM is better at learning information content of sequences in the same stretch of promoter region in *B.subtilis* compared *to E.coli*. Unlike certain *E.coli* training sets, HMMs trained well on all the sequence sets including the very last set i.e. 50_75 S_{50}(75), were. The most likely explanation for getting models to train well on *B.subtilis* promoter sets is probably due to the presence of several moderately conserved elements throughout most of *B.subtilis* promoter regions as suggested by Helmann, (1995). Equally, good results (low FP rates) were obtained with 101 bp test sequences with overall false positive rate being lower than what was obtained from *E.coli.*

UNIVERSITY of the
WESTERN CAPE

3.3.3 *Mycobacteria*

The same procedures used to carry out the promoter predictions after models were trained on *E.coli* and *B.subtilis* were applied to the study on *M.tuberculosis* sequence data. The major difference with regard to data between *M.tuberculosis* and the previous two is; other mycobacterial promoters were added to the collection of the original experimentally characterized *M.tuberculosis* promoter dataset. Also, the test dataset used consisted of only 34 promoters. It comprised of collection of mycobacterial promoters including those of *M.tuberculosis*. However, none of the test promoter datasets had the transcriptional start site experimentally characterized. Few have both –10 and –35 hexamers mapped experimentally but most had either the –10 or –35 hexamers experimentally characterized (refer to section on *M.tuberculosi*s Data). Thus the promoter regions used as test data were selected based on extrapolations from either one or both of known hexamer(s) in the sequence. The extrapolations based on the hexamer(s) definitely affected the results on *M.tuberculosis*, since the main established features of prokaryotic promoters are the two hexamers (-35 and –10) and the corresponding spacer region between them. However, this shortcoming may be reduced and may actually be less significant in the final prediction because a single prediction comprising the best in each category of inter-orf is selected instead of using a threshold value to categorize a sequence as a promoter or non-promoter as done in this chapter. Due to the adjustments necessary with regard to the data set of both training and test data, the results were not expected to be comparable to those accomplished on *E.coli* and *B.subtilis* datasets.

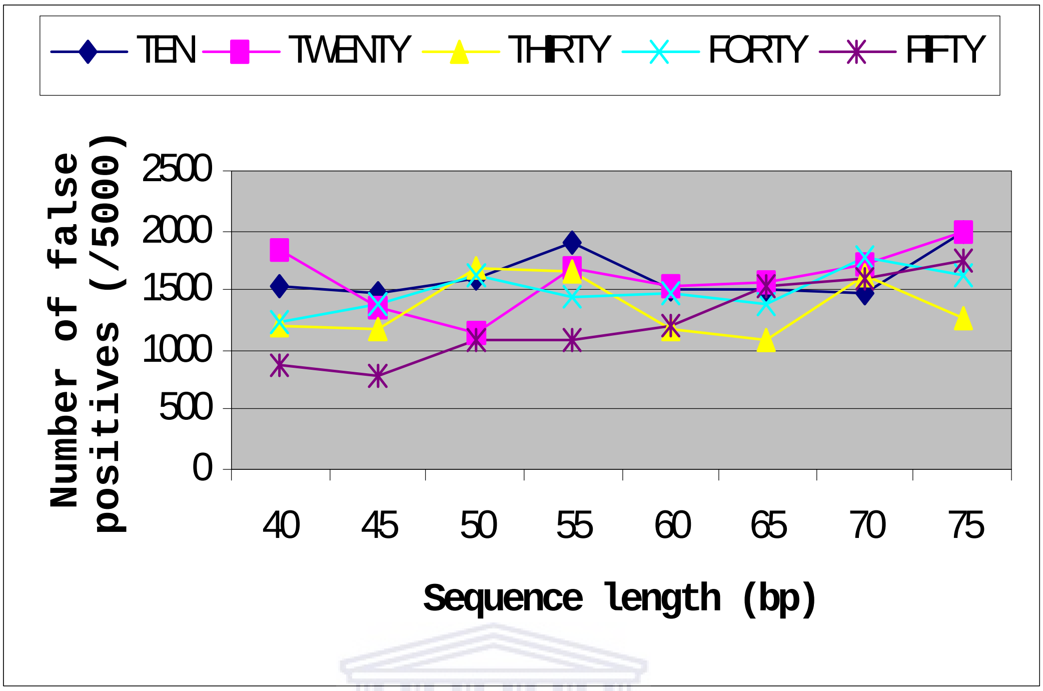
UNIVERSITY of the
WESTERN CAPE

Sets	1	2	3	4	5	Av.	%
10_40	1533	1544	1532	1529	1518	1531	30.6
10_45	1489	1496	1484	1479	1506	1490	29.8
10_50	1592	1583	1582	1582	1583	1584	31.7
10_55	1895	1952	1886	1930	1877	1908	38.2
10_60	1499	1535	1532	1520	1482	1513	30.3
10_65	1529	1476	1485	1527	146=3	1496	29.9
10_70	1462	1461	1455	1466	1497	1468	29.4
10_75	1948	1963	1975	1958	2024	1973	39.5
20_40	1858	1888	1805	1823	1812	1837	36.7
20_45	1358	1389	1307	1370	1366	1358	27.2
20_50	1164	1134	1128	1113	1125	1132	22.7
20_55	1656	1662	1672	1674	1712	1675	33.5
20_60	1533	1544	1552	1532	1543	1540	30.8
20_65	1549	1579	1577	1570	1596	1574	31.5
20_70	1699	1722	1697	1694	1700	1702	34.0
20_75	1968	2021	2003	2013	2010	2003	40.1

30_40	1243	1166	1200	1227	1214	1210	24.2
30_45	1193	1198	1143	1197	1205	1187	23.7
30_50	1664	1713	1695	1643	1679	1678	33.6
30_55	1640	1679	1642	1667	1697	1665	33.3
30_60	1164	1179	1180	1176	1196	1179	23.6
30_65	1079	1105	1084	1104	1109	1096	21.9
30_70	1673	1637	1641	1635	1597	1636	32.7
30_75	1241	1305	1283	1281	1291	1280	25.6
40_40	1294	1243	1260	1277	1180	1250	25.0
40_45	1381	1450	1366	1351	1421	1393	27.9
40_50	1624	1645	1628	1602	1632	1626	32.5
40_55	1441	1504	1395	1431	1464	1447	28.9
40_60	1499	1475	1487	1489	1468	1483	29.7
40_65	1397	1372	1370	1460	1397	1399	28.0
40_70	1801	1754	1796	1752	1721	1764	35.3
40_75	1575	1644	1637	1622	1614	1618	32.4
50_40	876	866	883	832	838	859	17.2
50_45	768	817	763	787	798	786	15.7
50_50	1111	1062	1081	1094	1072	1084	21.7
50_55	1071	1111	1070	1024	1106	1076	21.5
50_60	1184	1238	1180	1143	1223	1193	23.9
50_65	1554	1520	1526	1547	1540	1537	30.7
50_70	1621	1590	1616	1574	1560	1592	31.8
50_75	1723	1753	1783	1764	1727	1750	35.0

Table 3.8. False positive results obtained from trained HMM models on *M.tuberculosis* promoter data set on five thousand (5000) coding sequences. Promoter and non-promoter data set used in testing had the same fragment sizes as those of their corresponding models. For each non-promoter sequence that was tested, the average from five fragment sizes that corresponded to the model size was computed. The average scores for each model and the percent false positive scores are in the seventh and eight columns respectively. As in previous cases, threshold values that resulted in 90% TP were used.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Fig. 3.8. Individual trained HMM models with their corresponding false positive results on 5000 *Mycobacterial* coding sequences. Model 50_45 (fifty promoter sequences of fragment size 45 bp each) produced the best results (least number of false positives – 786). Models were tested on sequences having the same sequence length as those used in building the models. A cut-off score that produced 90 % (75/83) True positive (TP) was used to select the predicted promoters from non-predicted promoters.

As expected, false positive results are generally higher for all the three classes of test, figures 3.8, 3.9 and 3.10. Since many *M.tuberculosis* promoters are functional in other mycobacteria species (Mulder *et al*., 1997), one would expect the training data, though minimal, to be sufficient for promoter modeling on all the three methods (HMM, ANN and TFDA). The problem however, is most probably due to the test data. It is likely that, (a) the selected portions of the 30 promoters used for testing were not representative of the actual trained models or (b) there were not enough data for testing. Perhaps a distinctive feature, the absence of conserved –35 hexamer as proposed by Bashyam *et. al*., (1996) is being exposed in this HMMer study. Unfortunately, none of the mentioned hypothesis can be tested as there are not enough experimentally characterized promoters of *M.tuberculosis*. The results are still very encouraging; the best score for type *A* test being 15.7% false positive S_{50}(45). The results from sets of twenty and thirty were also very encouraging particularly those of S_{20}(50) and S_{30}(65). The pattern of results from same-as-model sequence to model *type C* test were generally similar to those of the previous two with bigger sequence sets producing better results than lesser sequence sets. Also, a direct correlation between fragment size and scores as is the case with *E.coli* and *B.subtilis* is not observed. The major noticeable difference between the pattern of results obtained on *M.tuberculosis* compared to those of *E.coli* and *B.subtilis* is the relative high false positive rated for different sets of test data.

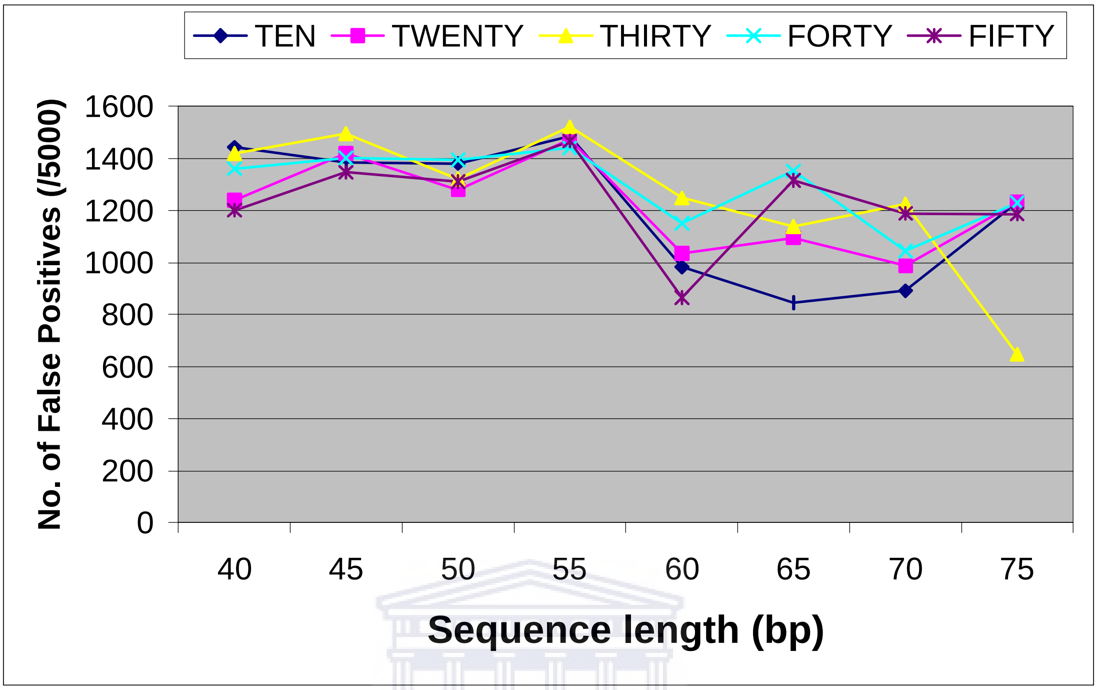
UNIVERSITY of the
WESTERN CAPE

SETS	1	2	3	4	5	Ave.	%
10_40	2715	2740	2742	2729	2749	1442	28.8

10_45	2669	2653	2667	2679	2669	1383.6	27.7
10_50	2663	2656	2674	2651	2662	1378.6	27.6
10_55	2810	2791	2804	2790	2797	1486.4	29.7
10_60	2153	2153	2169	2168	2164	980.8	19.6
10_65	2007	1995	2004	1984	2007	848	17.0
10_70	2056	2048	2036	2069	2049	890.4	17.8
10_75	2469	2462	2466	2460	2515	1230.6	24.6
20_40	2463	2489	2473	2465	2519	1239.2	24.8
20_45	2711	2711	2690	2719	2724	1418.8	28.4
20_50	2554	2548	2538	2509	2555	1280	25.6
20_55	2788	2784	2777	2775	2772	1471.6	29.4
20_60	2247	2215	2237	2228	2242	1034.4	20.7
20_65	2316	2288	2319	2300	2317	1094.8	21.9
20_70	2122	2183	2205	2138	2158	986.8	19.7
20_75	2459	2483	2473	2481	2466	1230.6	24.6
30_40	2718	2715	2705	2713	2708	1418.2	28.4
30_45	2800	2803	2S80	2806	2812	1494.2	29.9
30_50	2591	2585	2583	2583	2593	1318.8	26.4
30_55	2844	2839	2836	2840	2833	1519.6	30.4
30_60	2498	2509	2499	2493	2495	1249.2	25.0
30_65	2361	2353	2348	2362	2370	1136.6	22.7
30_70	2441	2454	2479	2494	2449	1225.2	24.5
30_75	1705	1757	1731	1728	1762	645.6	12.9
40_40	2646	2644	2635	2629	2643	1360.2	27.2
40_45	2681	2687	2680	2688	2701	1401.2	28.0
40_50	2677	2686	2676	2683	2680	1395	27.9
40_55	2729	2724	2738	2743	2739	1438.8	28.8
40_60	2371	2357	2394	2385	2359	1149	23.0
40_65	2610	2627	2635	2634	2615	1352.2	27.0
40_70	2211	2248	2270	2240	2216	1044.8	20.9
40_75	2443	2485	2464	2480	2467	1229.2	24.6
50_40	2443	2439	2438	2425	2444	1199.2	24.0
50_45	2602	2620	2614	2621	2619	1344.8	26.9
50_50	2574	2576	2572	2563	2584	1309	26.2
50_55	2780	2777	2767	2762	2774	1466	29.3
50_60	2036	2010	2019	2020	2025	864.8	17.3
50_65	2576	2581	2585	2587	2576	1315.8	26.3
50_70	2394	2407	2433	2446	2403	1187.8	23.8
50_75	2384	2435	2425	2412	2401	1184.6	23.7

UNIVERSITY of the
WESTERN CAPE

Table 3.9. False positive results of different trained models ranging from 10_40 to 50_75 on 5000 coding sequences of 75 bp fragment size each. Because the original sequence length of the test sequences are 101 bp, the average of five random sub fragments of 75 bp sequence length had to be used to give some credibility to the results. Sub fragments were generated by randomly selecting a position in the sequence that would make it possible to generate the 75 bp test sequence. The averages and percentage scores are shown on the seventh and eight columns respectively. On the left are the various models trained from respective sequence subsets.



UNIVERSITY of the
WESTERN CAPE

Fig. 3.9. Individual HMM models with corresponding false positive results on 5000 Mycobacterial coding sequences of 75 bp sequence length each. Each sequence's score was obtained by opening a window within the 75 bp sequence, which corresponded to the model size, and summing the results as the window was shifted 1 bp, fig. 3.1. Scores that resulted in 90% true positive from the 33 promoters were used as the cut-off score to distinguish between predicted promoters and non-promoters.

A

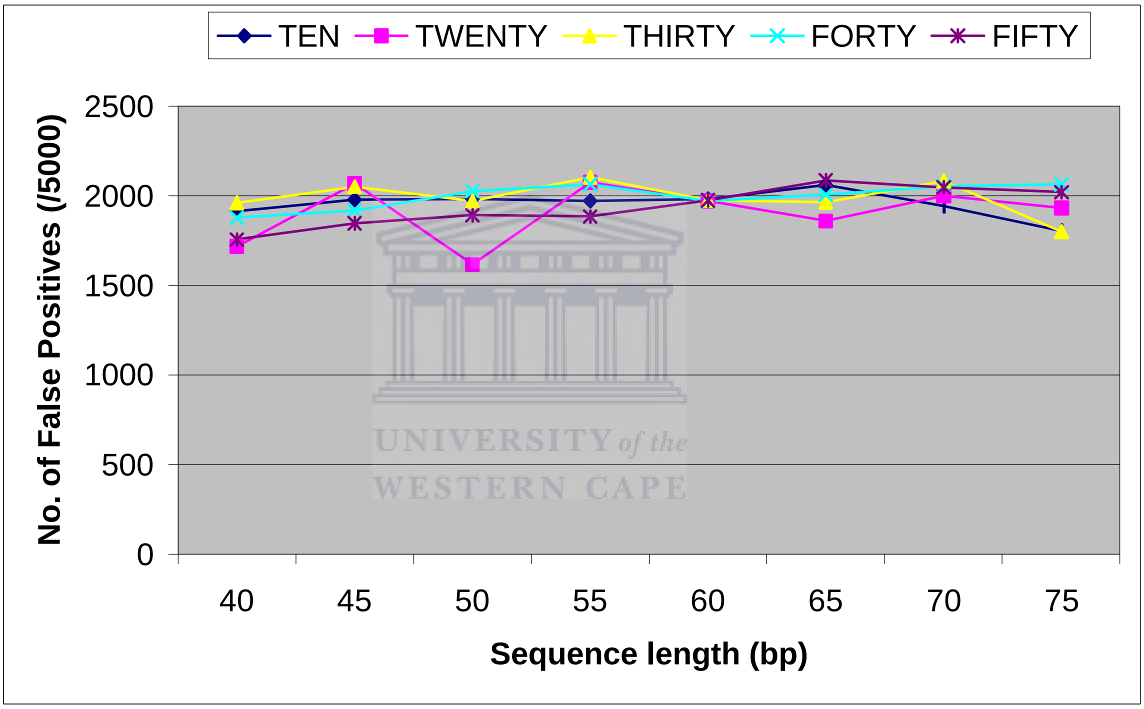
Sequence	Sequence Sets				
length	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	1915	1719	1961	1879	1756
45	1977	2069	2052	1918	1848
50	1983	1616	1972	2026	1891
55	1972	2077	2102	2064	1885
60	1983	1973	1976	1970	1974
65	2062	1859	1964	2006	2085
70	1942	1998	2084	2055	2047
75	1803	1934	1802	2065	2019

B

Sequence	Sequence Sets				
length	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	38.3	34.4	39.2	37.6	35.1
45	39.5	41.4	41.0	38.4	37.0
50	39.7	32.3	39.4	40.5	37.8
55	39.4	41.5	42.0	41.3	37.7
60	39.7	39.5	39.5	39.4	39.5
65	41.2	37.2	39.3	40.1	41.7
70	38.8	40.0	41.7	41.1	40.9
75	36.1	38.7	36.0	41.3	40.4

UNIVERSITY of the
WESTERN CAPE

Table 3.10A. Results obtained on 5000 coding sequences of 101 bp sequence length each using threshold values that resulted in 90% true positives. Table 3.10B is the percentage equivalent of the results obtained in table 3.10A.



UNIVERSITY of the WESTERN CAPE

Fig. 3.10. Individual HMM models with corresponding false positive results on five thousand (5000) Mycobacteria coding sequences of 101 bp fragment-size each. Each test sequence's score was obtained by opening a window within the 101 bp sequence, which corresponded to the model size, and summing the score as the

window was shifted 1 bp, fig. 3.1 (B). Cut-off scores that resulted in 90% true positives from the 33 promoters were used.

Model 50_45, 30_75 and 50_50 respectively produced the least number of false positives for the three classes of designed test as shown in figures 3.8, 3.9 and 3.10 respectively. Once again, overall observed pattern appear to be random. Nevertheless, some models especially those with fragment sizes ranging from 45 bp to 55 bp produced relatively fewer false positives compared to test sequences of larger fragment sizes.

An approach to promoter detection/prediction using HMMer with options of training and alignment has been used to study the information content of three prokaryotic promoter elements with reasonably satisfactory results. Since all promoter data available for three organisms namely *E.coli*, *B.subtilis* and Mycobacterium are aligned according to their transcriptional start sites (tss) with no attention paid to –10, and -35 and region in between the two hexamers, *hmmt* was the obvious choice of the HMMer package. This is because *hmmt* is able to train effectively on previously unaligned sequences. Training a HMM (*hmmt*) is an iterative process that seeks to maximize the probability that developed model(s) represent the example sequences. The model is not usually guaranteed to be the best model. So in order to obtain some reasonable degree of success with the models, many training sets were done for each sequence subset. The best HMM models (models with the least number of false positives) were selected for each sequence set. Results obtained for different organisms were similar in pattern. Some models produced very good results especially between *E.coli* (6.7% for 40_45) and *B.subtilis* (3.9% for 30_75). It is not a coincidence that models trained on promoter sequences 45 to 50 from the transcription start site (upstream) discriminated best against the coding sequences. This region S_(45) to (S_50) contains the canonical –10 and –35 boxes. A major observation from the three different organisms (*E.coli, B.subtilis and Mycobacterium)* in the study is; the lack

UNIVERSITY of the WESTERN CAPE

of a single subset of promoter sequences that consistently produced better results than other subsets. Each organism's promoter sequences produced unique results. This seems to suggest that, no particular sequence set can be earmarked as the set to produce the best results; the concept of the involvement of other transcriptional factors/accessories perhaps accounting for the inconsistent results obtained on all the test sequences. Each case of prokaryotic promoter sequence modeling using HMM detection must be treated and analyzed independently; the best model with respect to its prediction efficiency may then be used for the task of detecting promoter sequences from non-promoters for that particular organism. The results however suggest that, better results are obtainable from fragment sizes within the range of 45 bp-55 bp across all three organisms. This range might not necessary produce better results in the other two prediction methods.

UNIVERSITY of the
WESTERN CAPE

Chapter four

ANN studies on *E.coli, B.subtilis* and *Mycobacterium* promoters.

ABSTRACT

Three layered back-propagation networks were trained on various datasets of promoters and non-promoters from *E.coli, B.subtilis* and *Mycobacterium*. Promoter and non-promoter sequence datasets ranged from ten sequences of 40 bp fragment sizes (10_40) to fifty sequences of seventy-five bp fragment sizes (50_75). In most of the designed sequence subsets, (10_40 to 50_75), neural network models were successfully trained on the combined datasets of promoters and non-promoters. True positive (TP) prediction rates were set at 90% by manually selecting threshold scores. Promoter and non-promoter datasets ranging from 40 to 101 bp fragment sizes were tested with the trained neural network models. False positive (FP) rates as low as 6.6%, 6.7% and 13.9% were achieved for *E.coli, B.subtilis* and *M.tuberculosis* respectively on their respective datasets. The relatively high false positive (FP) rates for *M.tuberculosis* data may be attributed to 'not so clean' extrapolated sequence data as explained in the section on *M.tuberculosis* promoters (section 3.2.3.1).

UNIVERSITY of the
WESTERN CAPE

4.1. Introduction

Of the tools available to biologists involved in biological sequence analysis, perhaps the most promising is the use of artificial neural networks. This is because neural network offers a somewhat direct approach to the problem by direct learning of the information content in nucleotide sequences. There have already been studies by some researchers (Lukashin *et al*., 1989; O'Neil, 1990, 1992; Pedersen *et al*., 1995) in neural network approach to promoter detection. However, these studies were carried out on *E.coli* and most focussed on specific regions of promoters such as the –10, –35 and transcriptional start sites. By being selective in which regions to study, the authors probably missed an opportunity to learn some more information harbored in the set of promoter sequences used in their study. For example, regions as far as 25 bp after transcriptional start in *E.coli* (Lewin, 1997) have been found to affect activity of certain promoters. There is also an unanswered debate on which regions upstream of transcriptional start site contribute to promoter activity. Some researchers have postulated that, regions as far as 70 bp upstream of the tss may affect transcription (Lukashin *et al*., 1989). The preliminary objective of the work described in this chapter was to train neural network architectures on set of sequences covering about 100 bp of *E.coli, B.subtilis* and *M.tuberculosis.* This would hopefully help to find out which regions around transcriptional start sites harbor the strongest promoter signal(s). The study done in this chapter allowed every possible information with respect to nucleotides, that enables RNA polymerase to identify the promoter region to be detected. The entire section around promoter sequences covering ~100 bp were therefore trained and studied on various neural network structures. The approach, which is designed to pick best-trained models within a sequence frame of 101 bp by training on different fragment sizes from 45 bp to 101 bp of separate sequence sizes (10 to 50 sets) is quite different from what most other researchers have done to date. Best-trained models for each organism's promoter dataset would be used with other prediction methods to elucidate promoter sequences in entire genomes of three organisms namely *E.coli, B.subtilis* and *M.tuberculosis* (chapter six).

4.2. METHODS.

4.2.1.1. *E.coli* Promoter Sequences.

The same *E.coli* promoter sequences in section 3.2.1.1 were used in analysis in this chapter.

4.2.1.2. *E.coli* Non-Promoter Training Data

E.coli non-promoter sequences (~500 sequences) used for training were generated from *E.coli* coding sequences 'ecoli.ffn' (Genbank version 111). Sequence lengths of 101 bp were extracted from randomly selected coding sequences in the Genbank file 'ecoli.ffn'. Sequence subsets consisting of sets often (10) to fifty (50) sequences were randomly generated from the training sets (promoters and non-promoters). The subsets were further assorted into different subsets according to fragment size that ranged from 40 bp to 75 bp (table 3.1).

4.2.1.3. *E.coli* Non-Promoter Test Data.

Same as in section 3.2.12.

UNIVERSITY of the
WESTERN CAPE

4.2.2.1. *B.subtilis* Promoter Data.

Same as those used in section 3.2.2.1.

4.2.2.2. *B.subtilis* Non-Promoter Training Data

B.subtilis non-promoter sequences (500) were generated from *B.subtilis* coding sequence file 'bsub.ffn', obtained from Genbank (version 111). Sequence lengths of 101 bp were extracted from randomly selected coding sequences in the Genbank file 'bsub.ffn'. Data sets similar to those made for *E.coli* non-promoter sequences were created (10_40 to 50_75).

4.2.2.3. *B.subtilis* Non-promoter Test Data.

Same as in section 3.2.2.2.

4.2.3.1. *M.tuberculosis* Promoter Data

Same data set used in section 3.2.3.1 was used for the neural net training and testing of Mycobacterium promoter sequences.

4.2.3.2. *M.tuberculosis* Non-Promoter Training Data.

Five hundred (500) sequences were randomly extracted from *M.tuberculosis* coding sequence file 'Mtub.ffn' (Genbank version 111). Sequences ranging from ten (10) to fifty (50) were selected randomly from the 500 sequences and used to train neural network models/architectures to recognize non-promoter sequences.

4.2.3.3. *M.tuberculosis* Non-promoter Test sequences

The same *M.tuberculosis* data used for the HMM testing (section 3.2.3.2) was used in for neural network tests.

UNIVERSITY of the
WESTERN CAPE

4.2.3. ARTIFICIAL NEURAL NETWORK

4.2.3.1. ANN Software

The neural network package used this study is Artificial Neural Networks (ANN) freeware obtained from Nureka Artificial Neural Systems (ANS); http://www.bgif.no/nureka. The software comes in two packages, '*nn*' and '*xnn*'. *Nn* is a specification language for building artificial neural network simulators based on modular layered neural network models. With *nn*, the topology of such a network, along with training rules, activation functions, initializations and connectivity among others can be specified. The language consists of abstract high level statements that describe the topology, learning rules and input data of the network. *Nn* creates a C-function from these specifications, which, when called with the proper parameters, will execute the network on a user supplied dataset (patterns) and return the results as an output parameter. A network generated by the *nn* compiler can be run on train and recall mode. The *nn* compiler can also create an executable file directly, which is capable of performing both train and recall tasks of the network. *Xnn* is the graphical window interface component of *nn*.

4.2.3.2. ANN ARCHITECTURE

The architecture of the networks used is a feed forward network with three layers of neurons and trained using the back-propagation training rule. The software was compiled and executed on a UNIX workstation (SGI). Several versions of Backpropagation network were designed with the number of hidden neurons ranging from one (1) to seven (7) whilst the input layers varied from 160 (for 40 bp fragment size) to 300 (for 75 bp fragment size).

4.2.3.3. INPUT DATA

UNIVERSITY of the
WESTERN CAPE

DNA sequences were encoded into a string by using a coding scheme where each nucleotide is represented by four (4) binary digits: A = 0001, C= 0010, G = 0100 T= 1000 (Brunak, *et. al.*, 1991). It has been found that this leads to a significantly better performance than a more compact coding scheme (A = 00, T = 01, G = 10, C = 11) presumably due to the identical Hamming distances between the nucleotide encoding (Demeler and Zhou, 1991). However, the compact coding scheme problem can be eliminated by doubling the number of neurons in the middle layer. The output layer in all networks consisted of one neuron, which determined whether a given sequence was a promoter or not. Promoter sequences were trained to output a value of 0.9 compared to the value of 0.1 to non-promoter sequences. Neural network training was carried out on each promoter subset using same number of non-promoters (CDS) of same fragment sizes. The trained networks were then tested on different sets of promoter and non-promoter sequences of same fragment sizes as those used in training the individual models. In the other two tests, sequences of fragment sizes 75 and 101 bp were tested by opening windows within the test sequences which corresponded to the fragment size used in training the particular model and taking the cumulative score as the window is shifted one bp (fig.3.1).

UNIVERSITY of the WESTERN CAPE

4.3. RESULTS AND DISCUSSION

Sequence subsets of promoter and non-promoter have been trained on different structures of feed forward network. The objective has been to train these various different network structures to distinguish and predict promoter elements from non-promoter sequences. The subsets used in training consisted of sequence sizes ranging from ten (10) to fifty (50). These sequence sets were further subdivided into categories according to sizes that also ranged from 40 bp fragment size to 75. In training the various neural net models, a problem of defining optimal trained models arose. There was always the problem of models being either under-trained or over-trained. The problem became even more heightened as there were over 40 models be trained for each of the three organisms (*E.coli*, *B.subtilis* and

Mycobacteria). To overcome the problem, training of the various neural network models/architectures were terminated intermittently to test their predictability efficiency on the dataset of promoters and non-promoters. An epoch of 500 (100 per cycle) was chosen as the period to check how ‘well’ the networks(s) had trained by cross validating with test promoter and non-promoter sequences. Certain models went through training until they achieved the optimal level with respect to the ability to distinguish between the two set of sequences. Some of the trained models turned out to be optimal; others were not and had to be trained under constant supervision. Once a model has been trained successfully, a file with information on the training is created. This file enables results to be reproducible. Unlike the HMM study where profiles/models were trained entirely on promoter sequences, neural network training had to incorporate non-promoter sequences as well. Equal sizes (same sequence number and fragment sizes) of promoter and non-promoter sequences were used in each specific model’s training. Every test set up to 75 bp fragment size was carried out on five sets of sub fragments from the same sequence, that is, the first nucleotide in a sub fragment is selected from random position depending on the size of the test fragment. The strategy was adopted to minimize biases resulting from selecting particular sub fragments in the test sequence data.

UNIVERSITY of the
WESTERN CAPE

4.3.1. *E.coli*

Table 4.1, 4.2 and 4.3 show the results obtained from various models on test sequences of equal sizes as their corresponding trained models, 75 bp fragment sizes and 101 bp fragment sizes respectively.

UNIVERSITY of the
WESTERN CAPE

SETS	1	2	3	4	5	Av	%
10_40	1676	1660	1621	1641	1672	1654	33.1
10_45	1542	1586	1519	1532	1552	1546	30.9
10_50	1488	1517	1526	1494	1476	1500	30.0
10_55	1510	1484	1474	1504	1515	1497	29.9
10_60	1408	1398	1421	1410	1388	1405	28.1
10_65	1270	1249	1250	1260	1288	1263	25.3
10_70	1417	1386	1372	1419	1408	1400	28.0
10_75	1365	1367	1399	1398	1362	1378	27.6
20_40	1243	1243	1202	1239	1249	1235	24.7
20_45	1056	1098	1051	1062	1080	1069	21.4
20_50	1009	1030	1056	1047	1042	1037	20.7
20_55	1107	1089	1099	1065	1055	1083	21.7
20_60	1053	1039	1019	1067	990	1034	20.7
20_65	978	982	973	1003	1020	991	19.8

20_70	1107	1146	1164	1138	1142	1139	22.8
20_75	1049	1018	1047	1007	1030	1030	20.6
30_40	821	820	803	802	850	819	16.4
30_45	1039	1048	1050	1064	1077	1056	21.1
30_50	1287	1270	1283	1273	1304	1283	25.7
30_55	1110	1135	1130	1124	1161	1132	22.6
30_60	1469	1393	1398	1417	1389	1413	28.3
30_65	1774	1775	1804	1769	1780	1780	35.6
30_70	843	847	850	855	844	847	17.0
30_75	1175	1188	1144	1145	1167	1164	23.3
40_40	699	694	679	673	729	695	13.9
40_45	947	974	924	921	972	948	19.0
40_50	1037	1064	1058	1044	1067	1105	21.1
40_55	1100	1126	1077	1103	1107	1103	22.1
40_60	1370	1386	1443	1417	1380	1399	28.0
40_65	1187	1232	1215	1150	1173	1191	23.8
40_70	863	859	892	885	841	868	17.4
40_75	843	860	820	800	808	826	16.5
50_40	467	465	472	448	478	466	9.3
50_45	774	829	765	802	809	796	15.9
50_50	529	534	556	554	520	539	10.8
50_55	664	660	650	649	628	650	13.0
50_60	875	882	878	876	861	874	17.5
50_65	668	672	683	688	718	686	13.7
50_70	854	814	807	828	832	827	16.5
50_75	670	635	656	651	634	649	13.0

UNIVERSITY of the
WESTERN CAPE

Table 4.1. Five sets of sequence sub fragments were generated randomly from each test sequence and tested on models trained on promoters and non-promoters of same fragment sizes. Thus a model Ec40_50 which was trained on a set of 40 sequences of 50 bp fragment sizes were tested on sequences of 50 bp fragment sizes. The average results of the number of false positives from the five sets together with their percentage false positive are shown on the seventh and eighth column respectively.

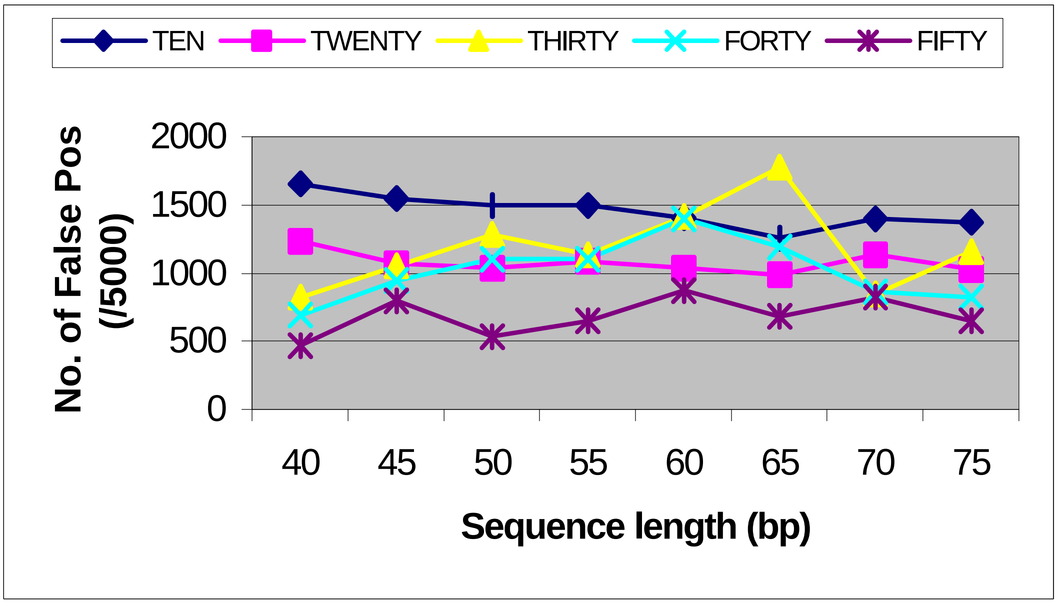


Figure 4.1. False positive prediction results (average) obtained from testing 5000 coding sequences using threshold values that resulted in 90% true positives for individual trained models. Test sequences had the same fragment sizes as the respective sequences used in training the models. Results from set fifty (50) produced relatively very good results with the best coming from model Ec50_40, a good low of 466 false positives out of 5000 test sequences (9.3%).

UNIVERSITY of the
WESTERN CAPE

SETS	1	2	3	4	5	Av.	%
10_40	1030	1032	1044	1045	1038	1037	20.8
10_45	919	894	922	917	899	910	18.2
10_50	947	942	930	907	934	932	18.6
10_55	831	880	869	858	847	857	17.1
10_60	960	941	970	959	953	956	19.1
10_65	620	648	665	636	628	639	12.8

10_70	1150	1088	1121	1125	1093	1115	22.3
10_75	1234	1228	1229	1193	1201	1217	24.3
20_40	538	534	544	552	543	542	10.8
20_45	455	459	454	441	455	453	9.0
20_50	545	570	558	554	564	558	11.0
20_55	425	434	448	446	435	438	8.8
20_60	409	390	392	397	386	395	7.9
20_65	465	444	470	472	456	461	9.2
20_70	733	749	769	731	770	750	15.2
20_75	1144	1137	1174	1167	1140	1152	23.1
30_40	467	480	471	476	473	473	9.5
30_45	473	471	472	476	483	475	9.5
30_50	455	466	453	462	449	457	9.1
30_55	437	425	446	433	428	434	8.7
30_60	485	516	480	485	498	493	9.9
30_65	441	427	470	450	436	445	8.9
30_70	543	547	529	528	522	539	10.7
30_75	1175	1214	1165	1200	1190	1189	23.8
40_40	538	542	545	530	530	537	10.7
40_45	597	574	591	575	594	586	11.7
40_50	551	544	551	526	527	540	10.8
40_55	575	572	534	551	544	555	11.1
40_60	437	451	448	465	442	449	9.0
40_65	376	389	409	410	409	399	8.0
40_70	575	567	577	586	557	572	11.4
40_75	905	961	951	949	928	939	18.8
50_40	393	388	382	403	401	393	7.9
50_45	406	408	406	405	398	404	8.1
50_50	450	445	451	461	436	449	9.0
50_55	498	494	505	498	517	502	10.0
50_60	623	599	626	618	620	617	12.3
50_65	501	487	497	501	501	497	9.9
50_70	592	609	623	594	577	599	12.0
50_75	869	868	897	898	840	874	17.5

UNIVERSITY of the
WESTERN CAPE

Table 4.2. The various neural net trained models and their corresponding results of false positives on 5000 coding sequences. Five sub fragments of 75 bp each were generated randomly from each test sequence as was done in the previous chapter used for testing. A threshold value that produced 90% true positive value on real promoter sequences was used in each case. The average results of the number of false positives from the five sets together with their percentage false positives are shown on the seventh and eighth column respectively.

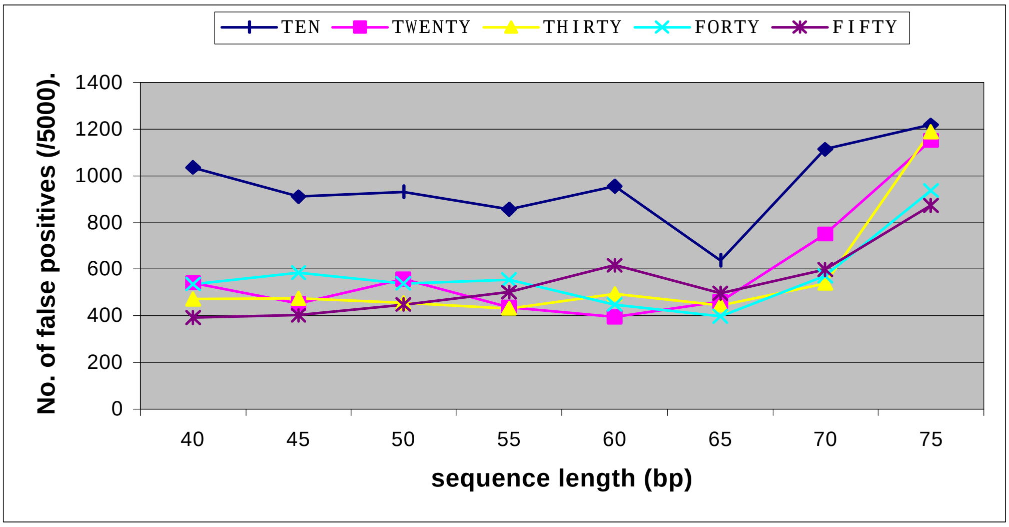


Figure 4.2. False positive prediction results (averages) obtained from testing 5000 coding sequences using threshold values for individual trained models that resulted in 90% true positives. Test sequences had fragment sizes of 75 bp. The average score from five data sets, created from each test of sequence (101 bp) was used Results from set Ec50_40 produced the best results of 393 (7.9%), though, an equally good results were obtained from the model Ec20_60 (395).

UNIVERSITY of the
WESTERN CAPE

A

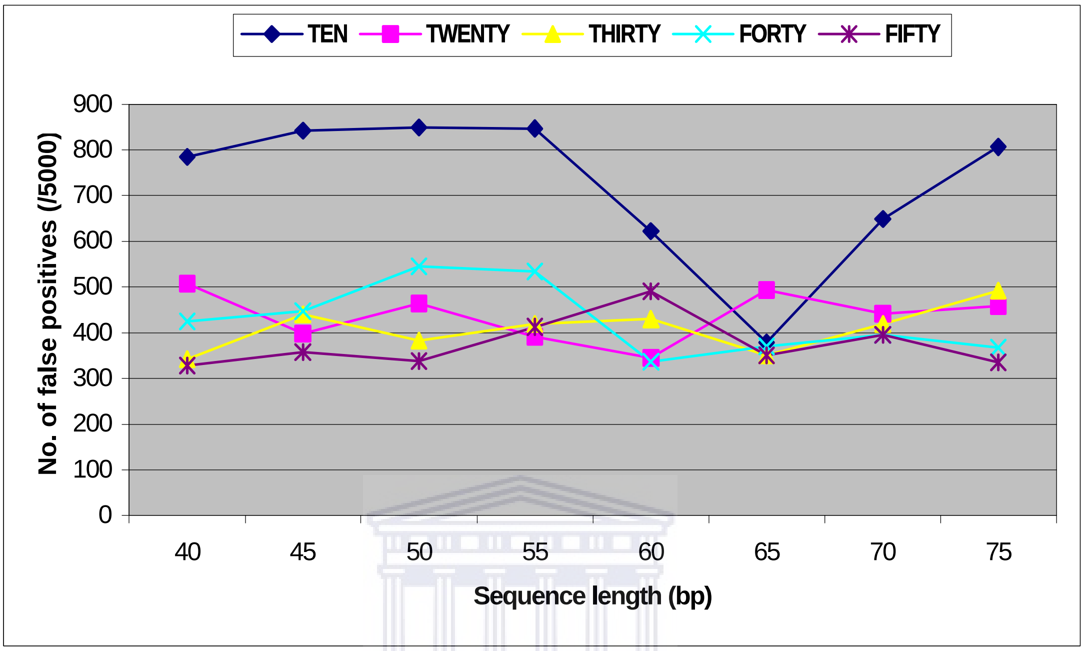
	TEN	*TWENTY*	*THIRTY*	*FORTY*	*FIFTY*
40	785	508	341	424	329
45	843	397	440	447	357
50	850	464	383	545	338
55	847	390	419	534	413
60	623	345	430	336	491
65	379	494	350	370	350
70	649	442	420	396	396
75	807	458	493	367	335

B

	TEN	*TWENTY*	*THIRTY*	*FORTY*	*FIFTY*
40	15.7	10.2	6.8	8.5	6.6
45	16.9	7.9	8.8	8.9	7.1
50	17.0	9.3	7.7	10.9	6.8
55	16.9	7.8	8.4	10.7	8.3
60	12.5	6.9	8.6	6.7	9.8
65	7.6	9.8	7.0	7.4	7.0
70	13.0	8.8	8.4	7.9	7.9
75	16.1	9.2	9.9	7.3	6.7

Table 4.3A. Results (false positives) obtained from various trained models on 5000 coding sequences. A threshold value that produced 90% true positive value on promoter sequences was used on the test set. Every sequence (101 bp) was tested by opening a window of size equivalent to the fragment sizes on which model was trained on, testing the model on the sequence and adding up the scores as window is shifted 1 bp. Table 4.3B consist of the results in table 4.3A expressed as percentages.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 4.3. False positive prediction results (averages) obtained from testing 5000 coding sequences using threshold values for individual trained models that resulted in 90% true positives for promoter sequences. The entire 101 bp fragment size of each sequence test set (both promoters and non-promoters) was used. Window sizes corresponding to model sizes were opened in test sequences and scores summed up as window was shifted 1 bp to the end of each sequence.

Analysis of the results revealed that, best prediction results were generally obtained from 40 bp up to 55 bp for almost all the models trained on the various sequence sets. As with the results obtained from HMM models (chapter three), larger sets resulted in overall better prediction results than smaller sets. Also, the overall results appeared to be best (least number of false positive) with typeC (test on sequence fragments of 101 bp) followed by *typeB* tests (test on sequence fragments of 75 bp); least false positives (percentage) being 6.6%, 7.9% and 9.3% respectively. The entire results were encouraging, especially on the dataset comprising fifty (50) sequences followed by forty sequences (figure 3.1). Again, as was the case with HMM, no obvious extended correlation was observed between fragment length and scores. It is worth drawing attention to the fact that, model trained on 50_40 consistently produced the best results in all the three test categories (*typeA*, *typeB and typeC*). Good prediction results were also obtained from the models trained on 20_60, 40_65 and 30_65.

The prediction results obtained from models tested on the entire 101 bp produced very similar results to those obtained on 75 bp sequence lengths, with better overall results (less false positives), ranging in numbers between 300 and 550. This result suggests that, with regard to neutral net training on *E.coli* promoter sequences, the longer the promoter region considered, the better the results. However, the results from the set of ten sequences though have relatively been poor in the previous cases, seem to be completely out of phase to the results from the other sets. Sequence subsets having 65 bp fragment sizes produced consistent results in this test category of 101 bp test datasets. Sixty five (65) bp fragments are therefore highly recommendable for training neural net models for prediction on 101 bp test datasets.

UNIVERSITY of the WESTERN CAPE

4.3.2. *B.subtilis*

Sequence subsets of *B.subtilis* promoter and non-promoter sequences have also been thoroughly trained on different architectures of Backpropagation network as done with *E.coli*. The objective, to train these various different network architectures to distinguish and predict promoter elements from non-promoter sequences. The same problem of over-training or under-training and identifying optimally trained models were encountered. The approach adopted in tackling the *E.coli* problem was also applied to this study. It involved use of procedural iteration to get the best-trained model for each promoter/non-promoter subset; by stopping the training process intermittently to test the predictability of the trained model on test set of promoters and non-promoters. Some of the models went through an automated training until they achieved optima, whilst others models had to be stopped from becoming over-trained. The results of the various models on test sequences of same size as those used to train the models, 75 bp sequence length and 101 bp are shown in tables 4.4, 4.5 and 4.6 respectively. The results are represented graphically in figures 4.4, 4.5 and 4.6.

UNIVERSITY of the
WESTERN CAPE

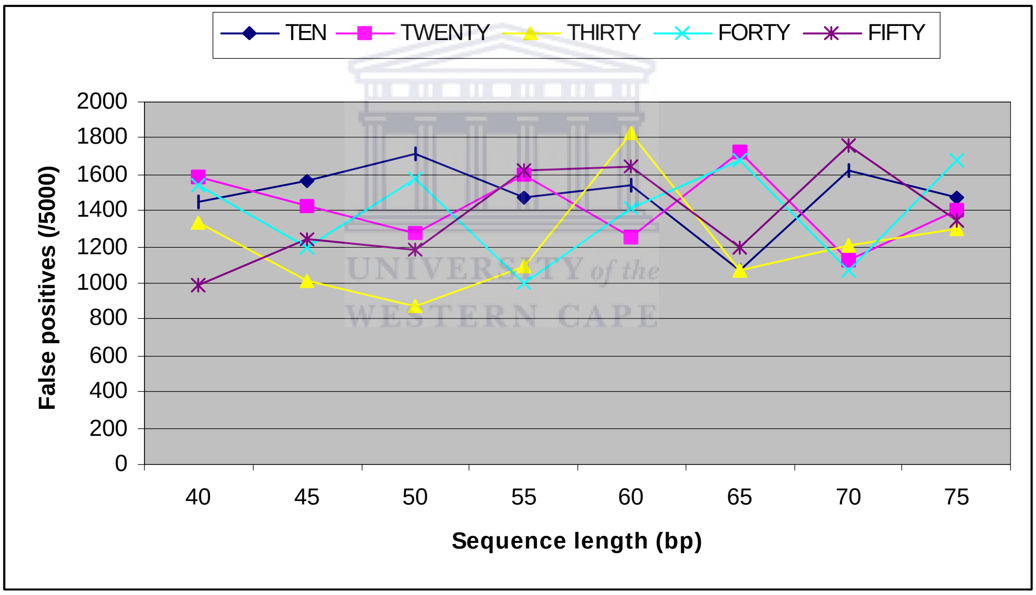
SETS	1	2	3	4	5	Av	%
10_40	1473	1447	1443	1442	1460	1453.0	29.1
10_45	1654	1650	1644	1648	1666	1652.4	33.0
10_50	1698	1716	1735	1696	1721	1713.2	34.3
10_55	1483	1470	1488	1441	1449	1466.2	29.3
10_60	1567	1559	1500	1535	1523	1536.8	30.7
10_65	1120	1055	1068	1054	1066	1072.6	21.5
10_70	1629	1606	1613	1632	1635	1623.0	32.5
10_75	1482	1469	1487	1477	1449	1472.8	29.5
20_40	1590	1605	1588	1560	1596	1587.8	31.8
20_45	1458	1400	1382	1439	1445	1424.8	28.5
20_50	1322	1287	1270	1258	1262	1279.8	25.6
20_55	1602	1609	1615	1590	1557	1594.6	31.9
20_60	1305	1250	1244	1246	1247	1258.4	25.2
20_65	1739	1712	1708	1741	1699	1719.8	34.4
20_70	1158	1148	1091	1107	1146	1130.0	22.6
20_75	1456	1391	1385	1405	1362	1399.8	28.0
30_40	1369	1314	1312	1319	1351	1333.0	26.7
30_45	1053	1028	998	997	986	1012.4	20.2
30_50	953	870	850	853	859	877.0	17.5
30_55	1140	1083	1054	1115	1066	1091.6	21.8
30_60	1847	1834	1809	1827	1805	1824.4	36.5
30_65	1107	1069	1027	1060	1059	1064.4	21.3
30_70	1225	1183	1189	1205	1215	1203.4	24.1
30_75	1338	1303	1285	1295	1256	1295.4	25.9
40_40	1569	1529	1540	1509	1546	1538.6	30.8
40_45	1252	1149	1181	1189	1185	1191.2	23.8
40_50	1636	1524	1580	1567	1588	1579.0	31.6
40_55	1071	985	989	966	997	1001.6	20.0
40_60	1409	1432	1403	1430	1373	1409.4	28.2
40_65	1670	1682	1659	1675	1698	1676.8	33.5
40_70	1103	1073	1052	1083	1059	1074.0	21.5
40_75	1684	1681	1683	1677	1687	1682.4	33.6
50_40	1041	978	986	952	1002	991.8	19.8
50_45	1283	1228	1239	1231	1240	1244.2	24.9

UNIVERSITY of the
WESTERN CAPE

50_50	1257	1162	1184	1150	1194	1189.4	23.8
50_55	1626	1585	1638	1618	1619	1617.2	32.3
50_60	1667	1645	1628	1611	1641	1638.4	32.8
50_65	1228	1177	1164	1212	1186	1193.4	23.9
50_70	1764	1780	1781	1745	1746	1763.2	35.3
50_75	1405	1332	1333	1318	1357	1349.0	27.0

Table 4.4. Results on five sets of sequence sub fragments generated randomly from each test sequence. These sub fragments were tested on models trained on promoters and non-promoters of same fragment size. Thus a model Bs40_50 trained on 40 sets of sequences of 50 bp fragment sizes were tested on 50 bp sequences. The average results of the number of false positives from the five sets together with their percentage false positive are shown on the seventh and eighth column respectively.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the WESTERN CAPE

Figure 4.4. Plot of false positive results (average) obtained from testing 5000 coding sequences using manually selected threshold values that resulted in 90% true positives for individual trained models. Test sequences had the same fragment sizes as the respective sequences used in training the models. Results from set

thirty (30) produced comparatively good results with the best coming from model composed of thirty sequences of fifty fragment sizes (Bs30_50).

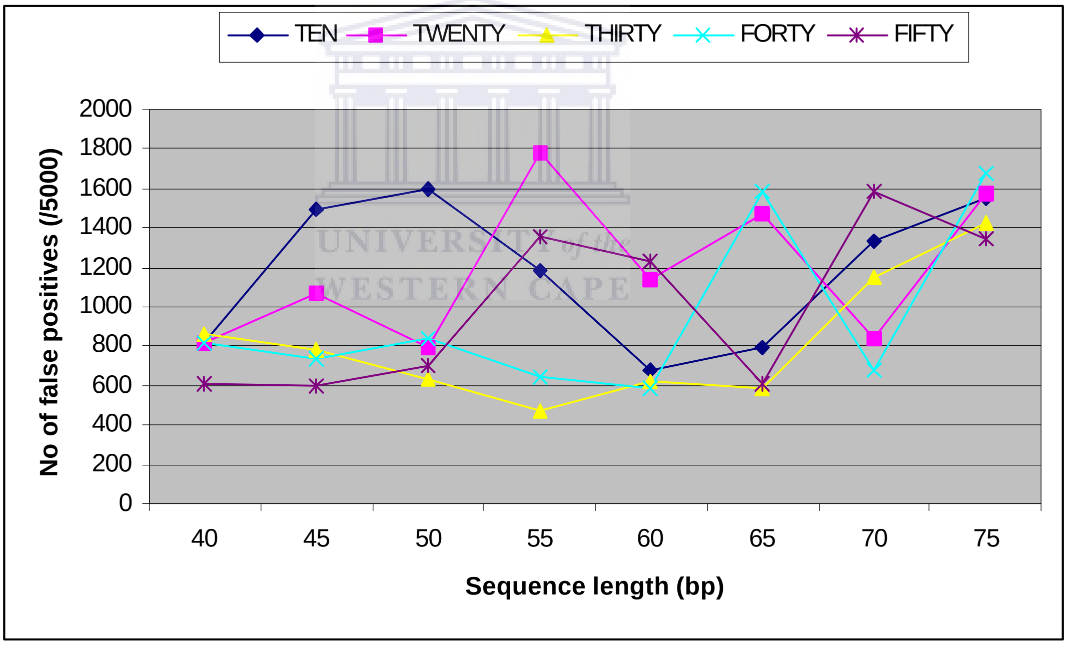
SETS	1	2	3	4	5	Av	%
10_40	781	814	795	890	810	818.0	16.4
10_45	1496	1489	1492	1529	1470	1495.2	29.9
10_50	1602	1577	1597	1625	1593	1598.8	32.0
10_55	1176	1161	1169	1256	1179	1188.2	23.8
10_60	657	652	673	773	656	682.2	13.6
10_65	758	759	778	910	785	798.0	16.0
10_70	1327	1323	1338	1405	1297	1338.0	26.8
10_75	1558	1556	1572	1552	1520	1551.6	31.0
20_40	792	804	793	901	802	818.4	16.4
20_45	1053	1039	1052	1172	1032	1069.6	21.4
20_50	763	762	780	872	778	791.0	15.8
20_55	1787	1794	1775	1798	1772	1785.2	35.7
20_60	1160	1100	1123	1149	1132	1132.8	22.7
20_65	1467	1463	1468	1519	1449	1473.2	29.5
20_70	798	844	833	915	825	843.0	16.9
20_75	1579	1584	1560	1613	1524	1572.0	31.4
30_40	840	825	836	956	842	859.8	17.2
30_45	766	775	752	946	759	799.6	16.0
30_50	588	599	598	756	594	627.0	12.5
30_55	446	443	438	573	431	466.2	9.3
30_60	598	611	599	728	593	625.8	12.5
30_65	566	558	540	708	555	585.4	11.7
30_70	1086	1153	1139	1251	1131	1152.0	23.0
30_75	1408	1435	1414	1454	1387	1419.6	28.4
40_40	808	799	782	916	771	815.2	16.3
40_45	705	708	701	835	705	730.8	14.6

UNIVERSITY of the
WESTERN CAPE

40_50	811	831	802	930	818	838.4	16.8
40_55	629	611	616	748	597	640.2	12.8
40_60	563	551	577	688	570	589.8	11.8
40_65	1576	1546	1583	1620	1579	1580.8	31.6
40_70	656	668	659	770	662	683.0	13.7
40_75	1684	1681	1683	1677	1687	1682.4	33.6
50_40	590	600	565	681	592	605.6	12.1
50_45	573	577	585	667	563	593.0	11.9
50_50	685	677	674	808	674	703.6	14.1
50_55	1303	1347	1344	1439	1334	1353.4	27.1
50_60	1220	1222	1235	1251	1231	1231.8	24.6
50_65	590	597	581	674	612	610.8	12.2
50_70	1592	1585	1559	1612	1575	1584.6	31.7
50_75	1318	1332	1333	1405	1357	1349.0	27.0

Table 4.5. Results (prediction) on various neural-net trained models and their corresponding results of false positives on 5000 coding sequences. Five sub fragments of 75 bp each were generated randomly from each test sequence and tested on the trained models. A threshold value that produced 90% true positive value on real promoter sequences was selected in each case. The average results of the number of false positives from the five sets together with their percentage false positives are shown on the seventh and eighth column respectively.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the WESTERN CAPE

Figure 4.5. False positive results (average) obtained from testing 5000 coding sequences using threshold values for individual trained models that resulted in 90% true positives. Test sequences had fragment sizes of 75 bp. The average score from

five data sets, created from each test of sequence (101 bp) was used. Results from model trained on thirty sequences of 55 bp sequence lengths (Bs30_55) produced the best results with regard to the number of false positives.

A

	TEN	TWENTY	THIRTY	FORTY	FIFTY
_40	744	640	560	608	486
_45	1663	696	643	628	447
_50	1181	638	499	786	699
_55	1143	1765	428	623	976
_60	655	1022	586	392	1156
_65	634	1493	659	1531	593
_70	1042	606	891	558	1364
_75	1310	1205	691	746	450

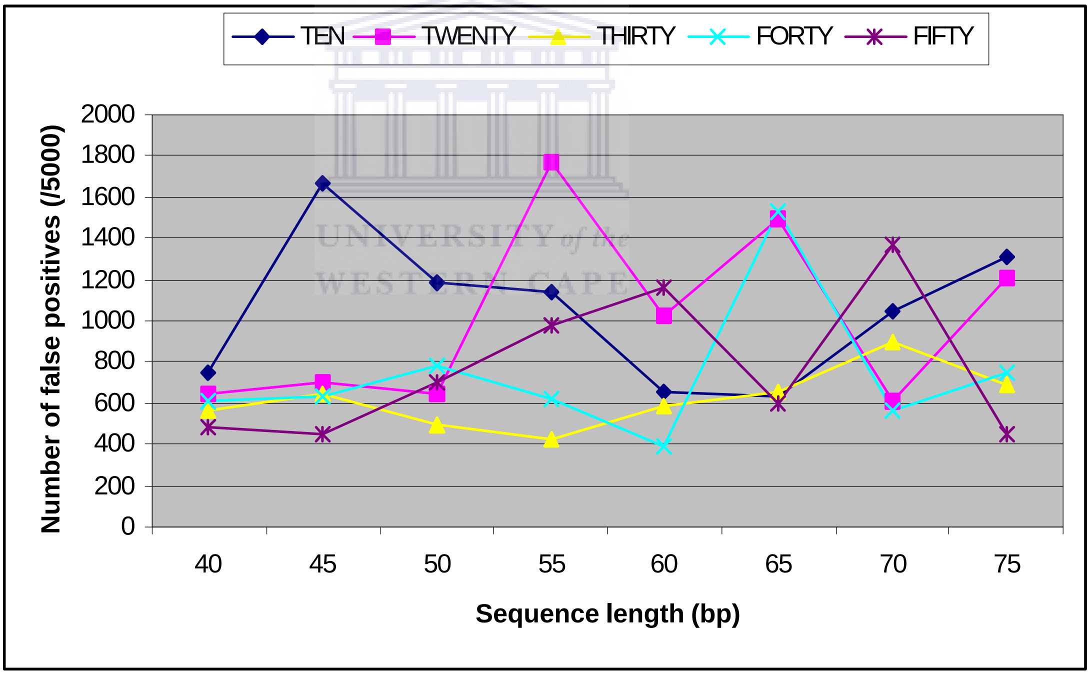
UNIVERSITY of the
WESTERN CAPE

B

	TEN	TWENTY	THIRTY	FORTY	FIFTY
_40	14.9	12.8	11.2	12.2	9.7
_45	33.3	13.9	12.9	12.6	8.9
_50	23.6	12.8	10.0	15.7	14.0
_55	22.9	35.3	8.6	12.5	19.5
_60	13.1	20.44	11.7	7.8	23.1
_65	12.9	29.9	13.2	30.6	11.9
_70	20.8	12.1	17.8	11.2	27.3
_75	26.2	24.1	13.8	14.9	9.0

Table 4.6A. Results (false positives) obtained from various trained models on 5000 coding sequences of *B.subtilis*. Table 4.6B is the equivalent of table 4.6A expressed

as percentages. Threshold values that produced 90% true positives on promoter test sequences were used. Every sequence (101 bp) was tested by opening a window of size equivalent to the fragment sizes on which model was trained on, testing the model on the sequence and adding up the scores as window is shifted 1 bp as described previously.



UNIVERSITY of the WESTERN CAPE

Figure 4.6. False positive results (average) obtained from testing five thousand (5000) coding sequences using threshold values for individual trained models that resulted in

90% true positives for promoter sequences. The entire 101 bp fragment size of each sequence test set (both promoters and non-promoters) was used. Window sizes that corresponded to the model sizes were opened in test sequences and scores summed up as window was shifted 1 bp to the end of each sequence.

The application of trained neural network models on *B.subtilis* promoters appear to follow a pattern similar to the results achieved with *E.coli* datasets; in that no obvious correlation is established between sequence size (number of sequences in the set) and prediction results. However, overall prediction patterns appear to be similar those obtained on *E.coli*. Larger (more number of sequences) sets generally produced better results compared to smaller sequence sets as generally reflected on sets of ten and twenty, figure 4.4. No sequence subset consistently produced bad results (high percentage of false positives) in all the three test categories. Sequence set of thirty produced the best (least number of false positives) results in the same-size-as-model category. Model trained on thirty (30) promoter sequences of fifty (50) bp fragment size (30_50) prediction resulted as the best score with 877 false positives out of five thousand (5000) sequences. Large fluctuations in prediction scores are observed for almost all the models. False positives range from 877 (17.5%) to an unacceptable high of 1824 (36.5%). Again, models producing very good results were those trained on sequences from the region with the canonical –35 and –10 hexamers.

UNIVERSITY of the
WESTERN CAPE

Results from the 75 bp category (sequences of 75 bp fragment sizes) revealed a different strength in model pattern. An impressive least score of 466 false positives out of five thousand (5000) test sequences is observed for model built and trained on 30_55. Model trained on thirty sequences of fifty bp (30_55), which performed best in the same size as model category also produced comparative results of (627/5000), fourth overall best with differences between the scores from the other two 40_60 and 30_60 being less than 40 sequences. The overall results, like those obtained for *E.coli* sequences of 75 bp were better than results on sequences having similar sequence length as the models. Still better results were obtained when entire 101 bp sequences were tested as compared to sequences of 75 bp fragment sizes. Results from testing entire 101 bp sequence fragments revealed yet another trained model 40_60 with a very impressive prediction score of 392 (7.8%). Prediction result obtained from model 30_55 (428) was still comparable to the best result of 392.

4.3.3. *M.tuberculosis*

UNIVERSITY of the
WESTERN CAPE

ANN was trained on mycobacterium promoter subset data as described in section 3.2.3.1 to develop various trained models capable of identifying *mycobacterium tuberculosis* promoters from non-promoters. The principle and rationale behind the experiment is the same as those used for *E.coli* and *B.subtilis* promoter and non-promoter datasets. Optimal training for each network model was achieved ‘manually’ for each model by stopping the training intermittently to test the prediction efficiency of the trained models on test promoters and non-promoters. Same problems of under-training and over-training in certain models came up. There were ~40 individual models to be trained. Tables 4.7, 4.8 and 4.9 show the various results obtained by testing sequences of same length as models, 75 bp fragment sizes and 101 bp fragment sizes respectively. Five thousand (5000) sequences from *M.tuberculosis* coding sequences were used for testing the non-promoters whereas only 34 promoter data were available for testing predictability on true positives.

SETS	1	2	3	4	5	Av.	%
10_40	1316	1280	1280	1286	1289	1290.2	25.8
10_45	2110	2097	2123	2103	2107	2108.0	42.2
10_50	1622	1600	1561	1565	1597	1589.0	31.8
10_55	1941	1961	1952	1965	2014	1966.6	39.3
10_60	2160	2150	2174	2162	2194	2168.0	43.4
10_65	1860	1887	1844	1881	1868	1868.0	37.4
10_70	1753	1750	1747	1766	1744	1752.0	35.0
10_75	1687	1732	1734	1750	1721	1724.8	34.5
20_40	1931	1916	1961	1943	1909	1932.0	38.6
20_45	2029	2023	2018	2002	2019	2018.2	40.4
20_50	1280	1269	1246	1250	1275	1264.0	25.3
20_55	1692	1694	1657	1683	1704	1686.0	33.7
20_60	2298	2292	2280	2292	2295	2291.4	45.8
20_65	1640	1622	1587	1649	1608	1621.2	32.4
20_70	1923	1899	1929	1903	1887	1908.2	38.2
20_75	1932	1863	1879	1890	1852	1883.2	37.7

UNIVERSITY of the
WESTERN CAPE

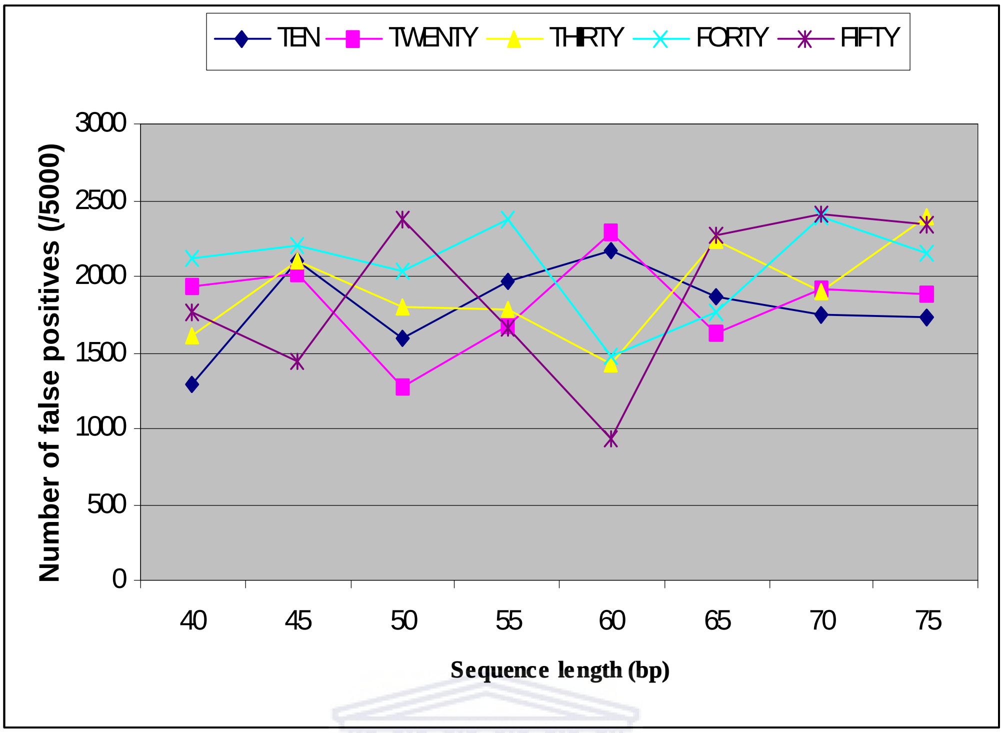
30_40	1584	1626	1600	1599	1626	1607.0	32.1
30_45	2082	2091	2110	2115	2134	2106.4	42.1
30_50	1817	1801	1777	1763	1796	1790.8	35.8
30_55	1769	1807	1792	1775	1779	1784.4	35.7
30_60	1443	1443	1439	1402	1414	1428.2	28.6
30_65	2260	2238	2223	2215	2220	2231.2	44.6
30_70	1913	1898	1915	1899	1869	1898.8	38.0
30_75	2398	2397	2399	2398	2398	2398.0	48.0
40_40	2119	2138	2103	2130	2130	2124.0	42.5
40_45	2217	2205	2217	2184	2190	2202.6	44.1
40_50	2043	2035	2040	2053	2020	2038.2	40.8
40_55	2381	2383	2377	2372	2377	2378.0	47.6
40_60	1504	1483	1486	1479	1464	1483.2	29.7
40_65	1725	1764	1787	1794	1743	1762.6	35.3
40_70	2394	2392	2395	2394	2393	2393.6	47.9
40_75	2129	2151	2130	2165	2162	2147.4	42.9
50_40	1751	1765	1777	1771	1742	1761.2	35.2
50_45	1429	1438	1469	1411	1499	1449.2	29.0
50_50	2364	2369	2382	2372	2380	2373.4	47.5
50_55	1631	1669	1666	1628	1677	1654.2	33.1
50_60	911	943	952	952	933	938.2	18.8
50_65	2251	2261	2285	2299	2273	2273.8	45.5
50_70	2399	2399	2399	2399	2399	2399.0	48.0
50_75	2328	2334	2333	2344	2340	2335.8	46.7

UNIVERSITY of the
WESTERN CAPE

Table 4.7. Results on five sets of sequence sub fragments generated randomly from each test sequence. These sub fragments were tested on models trained on promoters and non-promoters of same fragment size. Thus a model Mt40_50, trained on 40 sets of mycobacterium promoter sequences of 50 bp fragment sizes were tested on 50 bp sequences. Five thousand (5000) *mycobacterium*-coding sequences and 34 promoter sequences were used to test the models. Threshold values that resulted in 90% True Positive were selected from the promoter sequences and used as cut-off for the predictions. The average results of the number

of false positives from the five sets together with their percentage false positive are shown on the seventh and eighth column respectively.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 4.7. Plot of false positive results (average) obtained from testing 5000 mycobacterium coding sequences using manually selected threshold values that resulted in 90% true positives for individual trained models. Test sequences had the same fragment sizes as the respective sequences used in training the models. Best results (least number of false positives) came out of Mt50_60, model trained on fifty promoters of 60 bp fragment sizes. Thresholds from test promoter that resulted in 90% true positive were used to categorize ‘promoters’ from ‘non-promoters’.

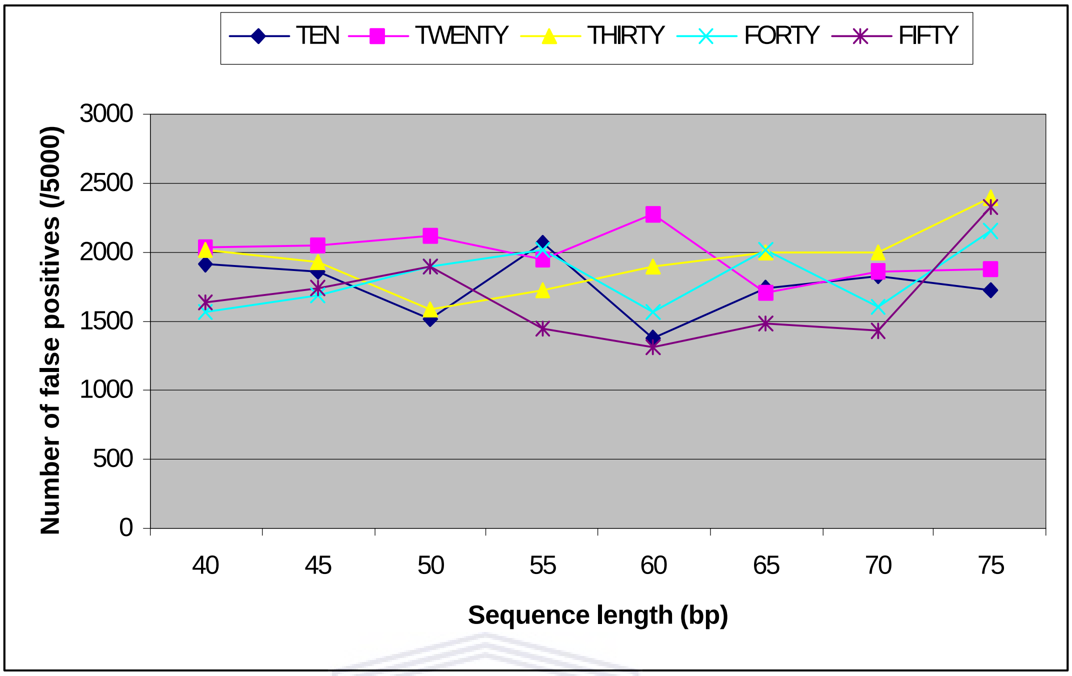
SETS	1	2	3	4	5	Av.	%
10_40	1903	1914	1917	1911	1917	1912.4	38.2
10_45	1850	1883	1884	1845	1884	1869.2	37.4
10_50	1504	1496	1518	1485	1540	1508.6	30.2

10_55	2057	2074	2062	2054	2074	2064.2	41.3
10_60	1379	1384	1390	1365	1407	1385.0	27.7
10_65	1763	1724	1750	1731	1746	1742.8	34.9
10_70	1804	1812	1839	1826	1826	1821.4	36.4
10_75	1750	1732	1734	1687	1721	1724.8	34.5
20_40	2042	2015	2026	2032	2038	2030.6	40.6
20_45	2060	2054	2047	2055	2047	2052.6	41.1
20_50	2127	2128	2121	2130	2116	2124.4	42.5
20_55	1955	1967	1968	1939	1936	1953.0	39.1
20_60	2263	2265	2267	2275	2273	2268.6	45.4
20_65	1731	1668	1743	1699	1711	1710.4	34.2
20_70	1896	1834	1872	1865	1875	1868.4	37.4
20_75	1890	1863	1879	1932	1852	1883.2	37.7
30_40	2026	2005	2013	2004	2013	2012.2	40.2
30_45	1935	1915	1939	1925	1938	1930.4	38.6
30_50	1591	1571	1594	1563	1598	1583.4	31.7
30_55	1717	1743	1691	1721	1743	1723.0	34.5
30_60	1868	1884	1892	1898	1907	1889.8	37.8
30_65	2008	1982	1979	2004	2006	1995.8	39.9
30_70	2003	2005	1973	1991	1988	1992.0	39.8
30_75	2398	2397	2399	2398	2398	2398.0	48.0
40_40	1580	1581	1570	1564	1585	1576.0	31.5
40_45	1697	1676	1684	1699	1673	1685.8	33.7
40_50	1903	1907	1875	1879	1904	1893.6	37.9
40_55	2017	2014	2032	2016	2014	2018.6	40.4
40_60	1593	1547	1568	1561	1590	1571.8	31.4
40_65	2015	2024	2031	2023	2020	2022.6	40.5
40_70	1621	1614	1597	1581	1640	1610.6	32.2
40_75	2165	2151	2130	2129	2162	2147.4	42.9
50_40	1638	1633	1635	1628	1651	1637.0	32.7
50_45	1737	1741	1733	1731	1745	1737.4	34.7
50_50	1900	1884	1918	1885	1906	1898.6	38.0
50_55	1470	1434	1428	1450	1479	1452.2	29.0
50_60	1309	1302	1297	1317	1309	1306.8	26.1
50_65	1477	1456	1507	1476	1478	1478.8	29.6
50_70	1412	1422	1457	1432	1471	1438.8	28.8
50_75	2344	2334	2333	2328	2340	2335.8	46.7

UNIVERSITY of the
WESTERN CAPE

Table 4.8. Results on various neural-net trained models and their corresponding results of false positives on 5000 mycobacterium coding sequences. Five sub fragments of 75 bp each were generated randomly from each test sequence and tested on the trained models. A threshold value that produced 90% true positive value on real promoter sequences was selected in each case. The average results of the number of false positives from the five sets together with their percentage false positives are shown on the seventh and eighth column respectively.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the WESTERN CAPE

Figure 4.8. False positive results (average) obtained from testing 5000 coding sequences of *M.tuberculosis* using threshold values for individual trained models that resulted in 90% true positives. Test sequences had fragment sizes of 75 bp. The average score from five data sets, created from each test of sequence (101 bp) was used. Results from model trained on fifty (50) sequences of sixty (60) bp sequence lengths (Mt50_60) produced the best results with regard to the number of false positives.

A

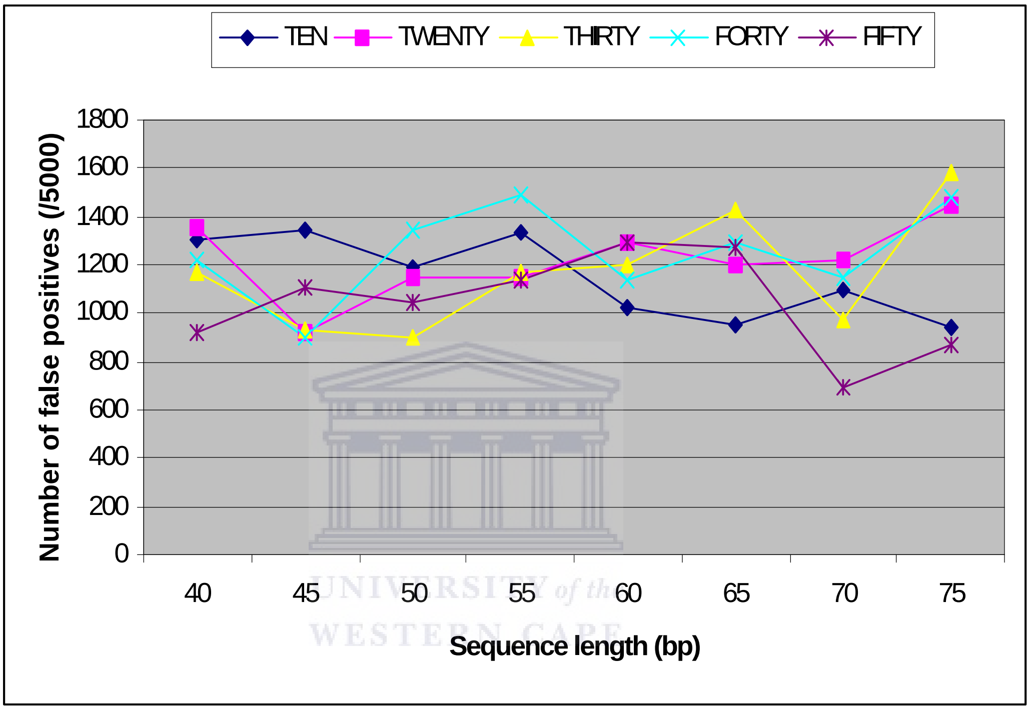
	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	1305	1354	1169	1223	918
45	1344	922	930	902	1111
50	1193	1153	897	1345	1045
55	1335	1146	1171	1488	1139
60	1029	1288	1203	1139	1289
65	955	1199	1429	1289	1269
70	1100	1220	969	1150	693
75	946	1451	1585	1484	864

B

	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	1305	1354	1169	1223	918
45	1344	922	930	902	1111
50	1193	1153	897	1345	1045
55	1335	1146	1171	1488	1139
60	1029	1288	1203	1139	1289
65	955	1199	1429	1289	1269
70	1100	1220	969	1150	693
75	946	1451	1585	1484	864

UNIVERSITY of the
WESTERN CAPE

Table 4.9. Results (false positives) obtained from various trained models on 5000 mycobacterium coding sequences. A threshold value that produced 90% true positive value on promoter sequences was used on the test set. Every sequence (101 bp) was tested by opening a window of size equivalent to the fragment sizes on which model was trained on, testing the model on the sequence and adding up the scores as window is shifted 1 bp.



UNIVERSITY of the WESTERN CAPE

Figure 4.9. False positive results (average) obtained from testing five thousand (5000) *M.tuberculosis* coding sequences using threshold values for individual trained models that resulted in 90% true positives for promoter sequences. The entire 101 bp fragment size of each sequence test set (both promoters and non-promoters) was used. Window sizes that corresponded to the model sizes were opened in test sequences and scores summed up as window was shifted 1 bp to the end of each sequence (figure 3.1).

Results obtained from testing the individual models on five thousand mycobacterium coding sequences and promoter sequences portray no correlation between size of training data and performance nor fragment size and performance, figure 4.7., suggesting little correlation or influence between training dataset and predictability of the network. This is similar to the results attained on *E.coli* and *B.subtilis* test data. Scores peak and dip reflecting consecutive results of better and worse in almost all the models. No outstanding predictive performance is observed in this study (tested sequence having the same data set as models). However, results from model Mt50_60 appear to be relatively good (18.8% false positives) though not comparable to the best results achieved for *E.coli* (9.3%). The overall results on models trained on Mycobacterium promoters and non-promoters (number of false positives) appear to be worse than the results obtained from *B.subtilis* and *E.coli*. Figure 4.8 depicts the plot of results obtained with individual models on fixed fragment sizes of 75 bp. The pattern observed on *E.coli* and *B.subtilis* with respect to relationship between increase in fragment size and the number of predicted false positives is observed here. Further increase from 75 bp to 101 bp, figure 4.9 results in lower false positive scores for almost all the models. Model 50_60 performed best in both study A and study B whereas 50_70 produced the best results for study C (101 bp test sequences).

A wide variety of multi-layered feed forward network structures have been developed and trained on promoter and non-promoter sequences of *E.coli, B.subtilis* and *Mycobacterium*. Inputs nodes from the various architecture ranged from 160 (40 bp region) to 300 (75 bp sequence region). The promoters subjected to the various network architectures were not classified into any categories especially with regard to which ones are transcribed from which sigma factors. Secondly, no attention was paid spacing classes (distance between –35 and –10) of the promoter. Results obtained from the study though not exceptional, are very promising and clearly demonstrate the ability of neural network to discriminate against certain variables when trained properly to do so. Other researchers had obtained better true positive results using neural network on specifically *E.coli* promoters (Demeler and Zhou, 1991 (98%); Lukashin *et al*., 1989 (96-98%); Mahadevan and Ghosh, 1994 (98%)). However, these researchers designed their neural networks to accommodate the already known information around the consensus hexamers (–35 and –10) and the spacing between the hexamers. On the other hand, O'Neil (1992), tried to use a generalized network to predict *E.coli* promoters of 16, 17 and 18 spacer classes and came up with a lesser true positive percentage of 60%. Aside from not incorporating possible dependencies and correlations of position specific bases into the study, the number of promoters used by the mentioned researchers for training far exceed the maximum of fifty (50) used in this study. Most of the trained networks in this study were used with a consistent degree of success in distinguishing promoter sequences from non-promoter sequences. In the study on the Mycobacterium promoter sequences in particular, the predicted results were disappointing compared to those of *E.coli* and *B.subtilis*. The disappointing results may be on Mycobacteria attributed partly to lack of enough information in the training sets rather than the inherent power of neural networks. This is particularly so in the case of *M.tuberculosis*, where the information used as test data constituted a collection of promoter data with experimentally undetermined transcriptional start sites. The relatively poor results may therefore be attributed to the threshold values (90% true positives) used as cut-off to classify test sequences. The overall performance in this case depends to a large extent on the true positives (actual promoters) used in the test. Because, a threshold value that automatically results in 90% TP is used as cut-off in the prediction. Neural network is no doubt a very powerful analytical tool and quite easy to use. The results clearly show the discriminatory ability of neural network if well trained. Coupled with the fact that it requires very little if any mathematical or programming skills, it is a very useful tool for studying promoter detection/prediction. However it does require patience and time to obtain optimally trained models, that is, models that do not over-generalize or under- generalize. As

UNIVERSITY of the
WESTERN CAPE

was the case with HMM, the best predictive models could be integrated and used on entire genomic sequences.

UNIVERSITY of the
WESTERN CAPE

Chapter five

Use of Statistical Analysis in study and prediction of *E.coli*, *B.subtilis* and *Mycobacterial* promoters

ABSTRACT

Statistical analyses of promoters and non-promoters belonging *to E.coli, B.subtilis and M.tuberculosis* datasets were performed. Statistical analysis performed included overall nucleotide composition, percent GC content, and dinucleotide/trinucleotide composition of the promoter/non-promoter dataset pairs of these organisms. Subtle but significant differences in nucleotide composition were observed between promoter and non-promoter sub datasets of equal sizes (equal number of promoters and non-promoters of same fragment sizes). These differences in composition were exploited to develop a prediction system named Triplet Frequency Distribution Analysis (TFDA). TFDA utilizes differences in trinucleotide composition of both promoters and non-promoters to produce a hash table of scores for each of the sixty-four possible triplets. Results of TFDA on promoter prediction were very comparable to those obtained from ANN (Artificial Neural Net) and HMM (Hidden Markov Model). TFDA produced true positive (TP) results of 90% and best false positive prediction results of ~5.9%, ~5.9% and ~20.4 for *E.coli, B.subtilis and M.tuberculosis* datasets respectively. The high false positive rate obtained on *M.tuberculosis* may be attributed to the minimal size of Mycobacteria promoter test data. Our analysis reveals that this statistical method predicts promoter sequences effectively with minimal errors when compared to other approaches such as HMM and NN.

UNIVERSITY of the
WESTERN CAPE

5.1. Introduction

Alignment of sequences having the same or similar functions have enabled researchers to identify certain novel consensus features in these sequences. In some cases, it has resulted in identification of the function of previously unknown sequences. Thus sequence alignment has been the backbone of scientific approach to elucidate function(s) of previously unknown sequences. A typical example is the identification of the canonical –10 and –35 hexamers of *E.coli* promoters (Hawley and McClure, 1983; Harley and Reynolds, 1987; Lisser and Margalit, 1993). At the backbone of sequence analysis via alignment, is the composition of DNA in the sequence strings. The very fact existence of codon usage preferences in organisms emphasizes the significance of importance of skewed nucleotide composition. These differences in various regions of the entire genome have been exploited in gene prediction/finding. (Krogh, 2000; Rees *et al*., 2000; Kulp *et al*., 1997; Shmatkov *et al*., 1999). A coding region of even the AT-rich *E.coli*, would not be expected to contain many aggregates of A's and T's; in case some form of mutation results in a stop codon (TAA, TGA and TAG) in the middle of the string. Likewise, one does not expect many strings of C and G's in promoter regions of even the GC-rich *M.tuberculosis* as that would result in more energy to open up the helix in the process of forming an 'open-enzyme-promoter complex. Within the same organism (*E.coli*), statistical analysis of nucleotide composition in the genome has enabled the recognition of certain genes as `acquired genes' (Medigue *et al*., 1991; Munoz, 1998). 'Acquired genes' are genes thought to acquired later through evolution by horizontal gene transfer. Statistical analysis has already been applied to promoter detection (Cardon and Stormo, 1991; Horton and Kaneshia, 1992; Ozoline *et al*., 1997; Besemer and Borodovsky, 1999), though none of these researchers dealt directly with dinucleotide and trinucleotide composition of the datasets of the organism, which in all cases was *E.coli*. Nucleotide composition analysis in the

UNIVERSITY of the WESTERN CAPE

form of TFDA was used to carry out detailed analysis of nucleotides in both promoters and non-promoters of *E.coli, B.subtilis* and Mycobacterium. The outcome of the analysis led to a prediction system being built on the information gained from the analysis. This prediction system has been employed with some degree of success in predicting promoter sequences from the three organisms.

5.2. METHODS.

5.2.1.1. *E.coli* Promoter Sequences.

As in section 3.2.1.1 were used in this chapter.

5.2.1.2. *E.coli* Non-Promoter Training Data.

Same as in section 4.2.1.2.

5.2.1.3. *E.coli* Non-Promoter Data

The same *E.coli* non-promoter sequences in section 4.2.1.3.

UNIVERSITY of the
WESTERN CAPE

5.2.2.1. *B.subtilis* Promoter Data.

Same as those used in section 3.2.3.1.

5.2.2.2. *B.subtilis* Non-Promoter Training Data

Same as in section 4.2.2.2.

5.2.2.3. *B.subtilis* Non-Promoter Data

As in section 4.2.2.3.

.

5.2.3.1. *M.tuberculosis* Promoters

Same data set used in section 3.2.3.1 was used for the neural net training and testing of Mycobacterium promoter sequences.

5.2.3.2. *M.tuberculosis* Non-promoter Training Data

As in section 4.2.3.2.

5.2.3.3. *M.tuberculosis* Non-Promoter sequences

As in section 4.2.3.3.

5.3. Triplet Frequency Distribution Analysis (TFDA).

5.3.1. Production of Promoter/Non-promoter Hash Tables.

Promoter and non-promoter sequences were divided into sets and subsets as described in the methodology section of chapter three. Each promoter and non-promoter subsets (same number of sequences and fragment sizes) of the three organisms was analyzed for triplets in nucleotide composition. The triplet frequency of each promoter non-promoter dataset pair was obtained by the following procedure:

UNIVERSITY of the
WESTERN CAPE

(a) A three bp size window is opened from the first nucleotide in each sequence in the sequence set.

(b) An inventory of all the triplets in each sequence in the set was taken as the window is moved by one base pair (1 bp) to the end of the entire sequence.

(c) Similar triplets were grouped and counted to obtain the numbers present for all 64 possible triplets in the set.

(d) Actual triplet frequency in a particular sequence set was obtained by using the following formula:

$$f_{triplet} = \frac{(N_t)(4^3)}{M} \qquad 5.1.$$

Where N_t represent the number of times a particular triplet occurs in the sequence, M is the total number of nucleotides in the entire sequence set and $f_{triplet}$ denotes the actual frequency of the triplet in the set. Hash tables were created by subtracting the frequency of a particular triplet in non-promoter set from the corresponding frequency of the same triplet in the promoter dataset. Thus, triplets more prevalent in promoter sequences will have relatively high positive scores compared to triplets more prevalent in non-promoter sets (they will actually have negative scores) in the hash table, figure 5.1.

5.3.2. Scoring System.

Test sequences are appraised by: Opening a three base pair window, shifting the 3 bp window by a base pair at a time and adding up all the corresponding hash table values of the triplets in the sequence. Thus to compare sequences, the sequences in question have to be of the same sequence lengths. The higher the score on the test sequence the better the chances of the sequence in question having promoter function. In all the study cases, ranking was used to select a threshold value that would result in 90% true positive for the known promoter test sequences. Test sequences are then assessed using the selected threshold value. All computations and analysis of sequence data were done on a SGI (irix 6.3) workstation. Computational codes were written in C, C++ and Perl programming languages.

5.4. Results and Discussion

5.4.1. Nucleotide Composition (Promoters and Non-promoters)

Nucleotide composition and percentage GC content of promoters and non-promoters of the three organisms were studied by, carrying out statistical analysis on the same number of sequences and same fragment size for each dataset. Table 5.1 shows the results of nucleotide composition analysis on the three organisms.

UNIVERSITY of the
WESTERN CAPE

Results shown in table 5 are illustrated graphically in figure 5.1. In all the three organisms, promoter regions appear A/T rich compared to non-promoter regions, which rather appear to be *G/C* -rich.

UNIVERSITY of the
WESTERN CAPE

	E.coli		*B.subtilis*		*Mycobacterium*	
	NP	P	NP	P	NP	P
A	24.4	27.9	29.7	33.8	19.2	20.9
C	24.7	21.8	19.3	14.8	32.1	28.4
G	27.6	20.5	24.2	18.8	30.1	31.0
T	23.3	29.7	26.9	32.6	18.6	19.7
%GC	52.3	42.3	43.5	33.6	62.2	59.4

Table 5.1. Nucleotide composition of Promoters (P) and Non-promoters (NP) *of E.coli, B.subtilis* and *Mycobacterium*. Also included is the percentage composition

of GC content. Equal lengths of sequences were analyzed to obtain the above results.

UNIVERSITY of the
WESTERN CAPE

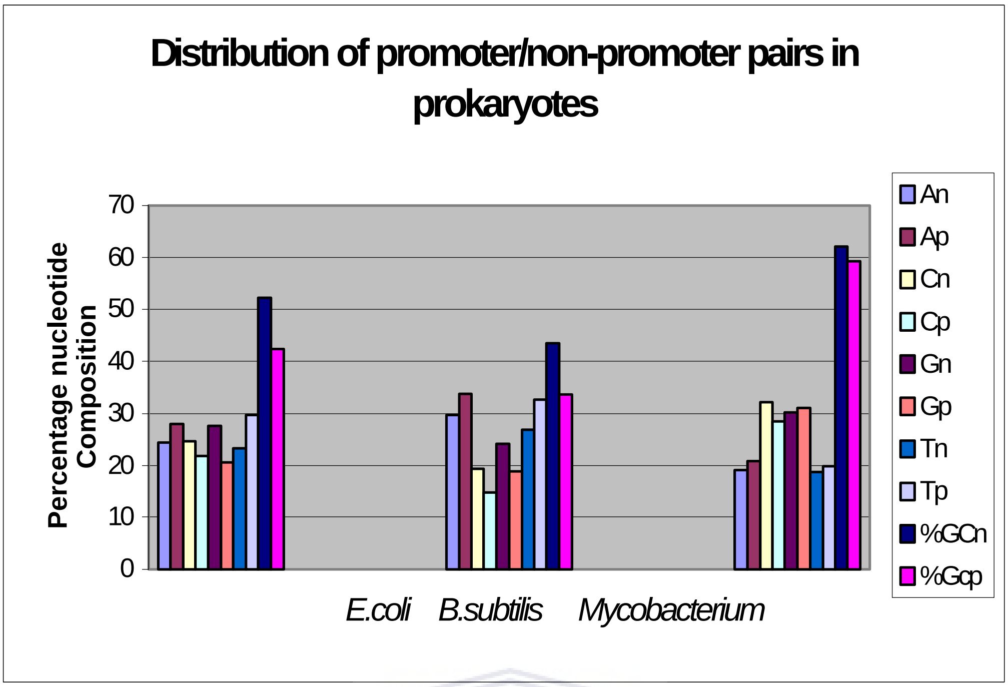


Figure 5.1. Percent nucleotide composition of promoter (Xp) and non-promoter sequences (Xn) obtained on *E.coli*, *B.subtilis* and *Mycobacterium* sequences. Sequences analyzed did not include the complements. Highest GC scores are observed for Mycobacterium sequences whilst least GC content is observed for *B.subtilis*.

UNIVERSITY of the
WESTERN CAPE

The organism with the highest GC content in both promoters and non-promoters is Mycobacterium. That confirms what has already been established (Kvasnikov *et al*., 1978; Danchin, 1997; Raghavan *et al*., 2000). Among the three organisms, mycobacterium is the most GC rich organism. Also, in all the three organisms, the percent *A* and *T* composition in promoters are relatively higher than those in their respective non-promoter dataset. The opposite is true for non-promoters, where the percent composition of G and C in the non-promoter data set is higher than found in their corresponding promoter data. Though both *B.subtilis* and *E.coli* are relatively *AT* rich organisms with respect to their nucleotide composition, *B.subtilis* has the higher A/T content in both its promoter and non-promoter data compared to *E.coli*.

5.4.2 DINUCLEOTIDE COMPOSITION (Promoter/Non-Promoter)

Table 5.2 shows the results obtained by carrying out dinucleotide analysis on the promoter (P) and non-promoter (NP) datasets of *E.coli, B.subtilis* and the *Mycobacterium* data. Dinucleotides AA, AT and TT (figure 5.2) appear to have very significant differences in the percentage composition in their promoter and non-promoter sequences with the higher score reflecting on the promoter sequences. On the other hand, the dinucleotides CG, GC and GG stand out in the percentage composition in both promoters and non-promoters of all the three organisms, though in most cases, it is the non-promoter data sets that have higher values of the dinucleotides.

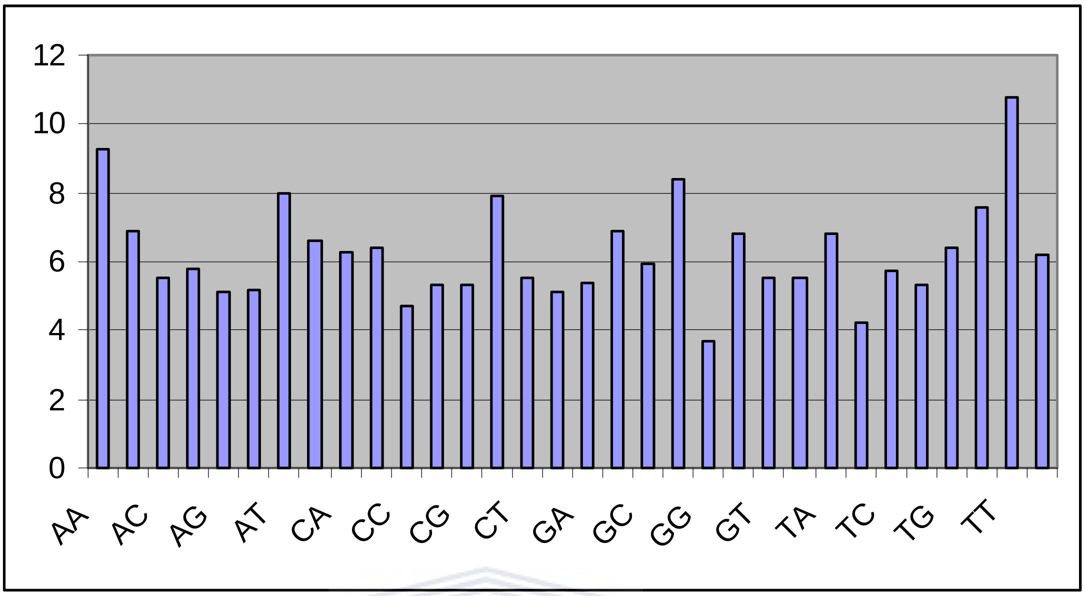
UNIVERSITY of the
WESTERN CAPE

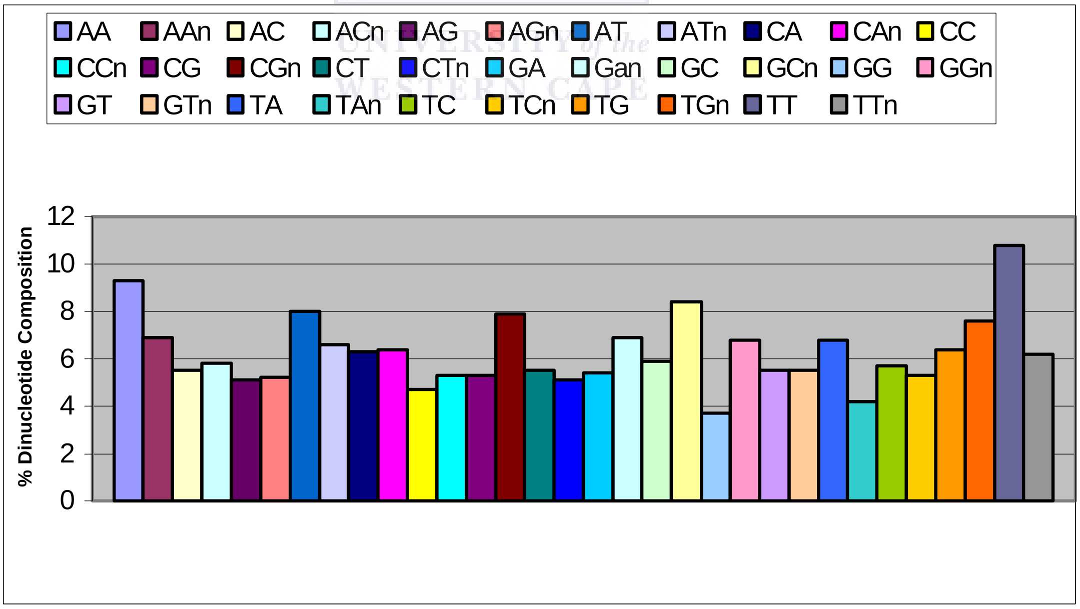
	E.coli		***B.subtilis***		**Mycobacterium**	
	P	**NP**	**P**	**NP**	**P**	**NP**
AA	9.3	6.9	13.0	10.4	4.4	4.1
AC	5.5	5.8	4.9	4.9	6.3	7.0
AG	5.1	5.2	5.7	6.1	5.6	3.9
AT	8.0	6.6	10.2	8.2	4.5	4.2
CA	6.3	6.4	5.6	6.0	5.8	6.9
CC	4.7	5.3	2.6	3.3	8.1	8.0
CG	5.3	7.9	2.5	4.8	9.8	11.9
CT	5.5	5.1	4.1	5.2	4.6	5.1
GA	5.4	6.9	6.6	7.7	7.4	6.2
GC	5.9	8.4	2.8	5.5	8.2	10.2
GG	3.7	6.8	4.3	5.7	9.0	8.2
GT	5.5	5.5	5.1	5.3	6.6	5.6
TA	6.8	4.2	8.7	5.5	3.3	1.8
TC	5.7	5.3	4.6	5.6	5.6	7.0
TG	6.4	7.6	6.3	7.7	6.7	6.1
TT	10.8	6.2	13.0	8.1	4.1	3.8

Table 5.2. Results obtained by computing the dinucleotide composition of large data sets (+80 sequences per data) of promoters (P) and non-promoters (NP) of *E.coli, B.subtilis* and *Mycobacterium*. Promoter and Non-promoter data for both *E.coli* and *B.subtilis* consisted of 8000 nucleotides each whilst Mycobacterium promoter and non-promoter datasets constituted 5000 nucleotides each. Outstanding differences in composition of between promoters and non-promoters of certain dinucleotides are observed in all three organisms. They include TT, AA, AT and in *E.coli* and *B.subtilis*, and GC and CG in mycobacterium.

UNIVERSITY of the
WESTERN CAPE

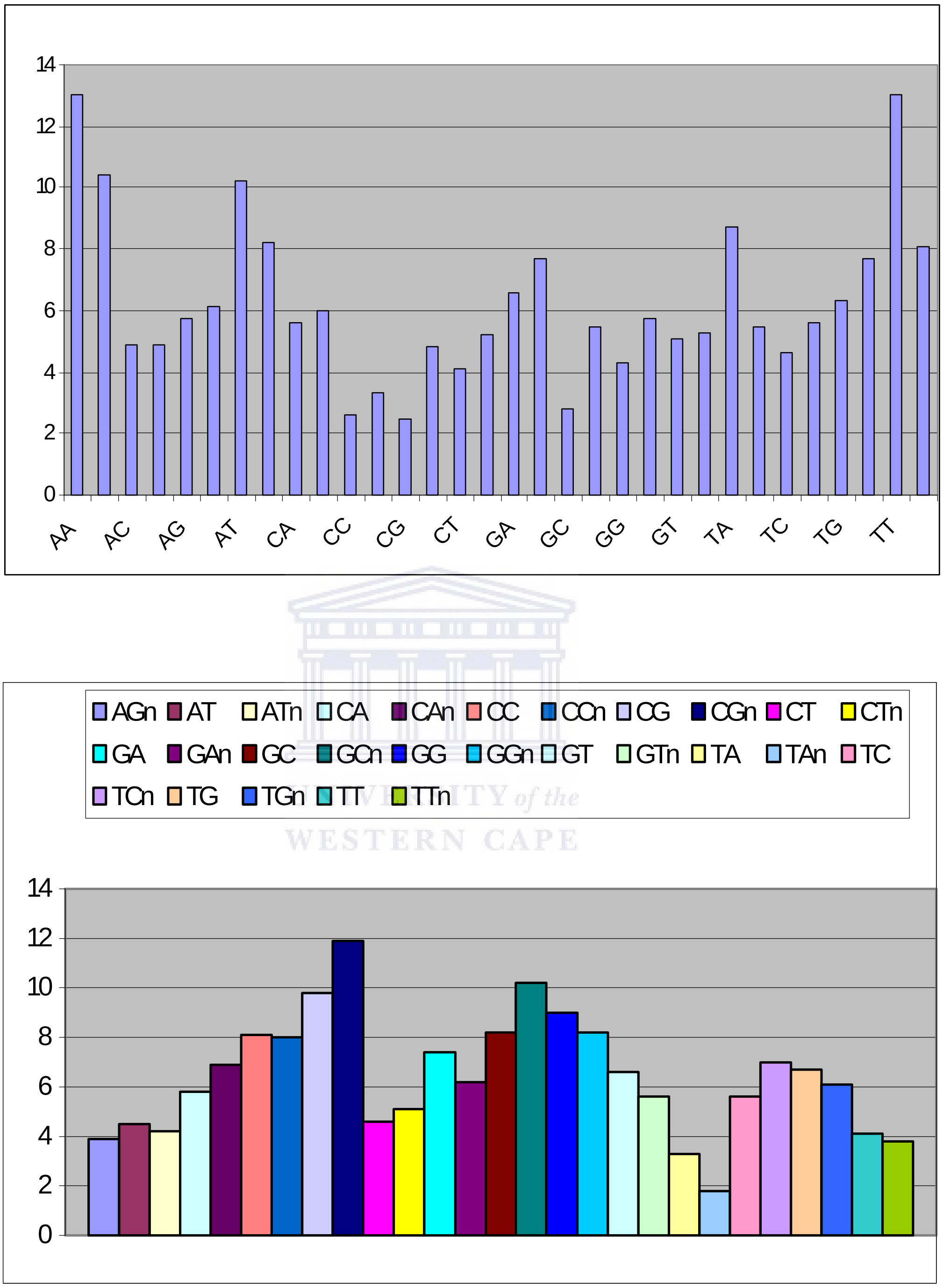
A





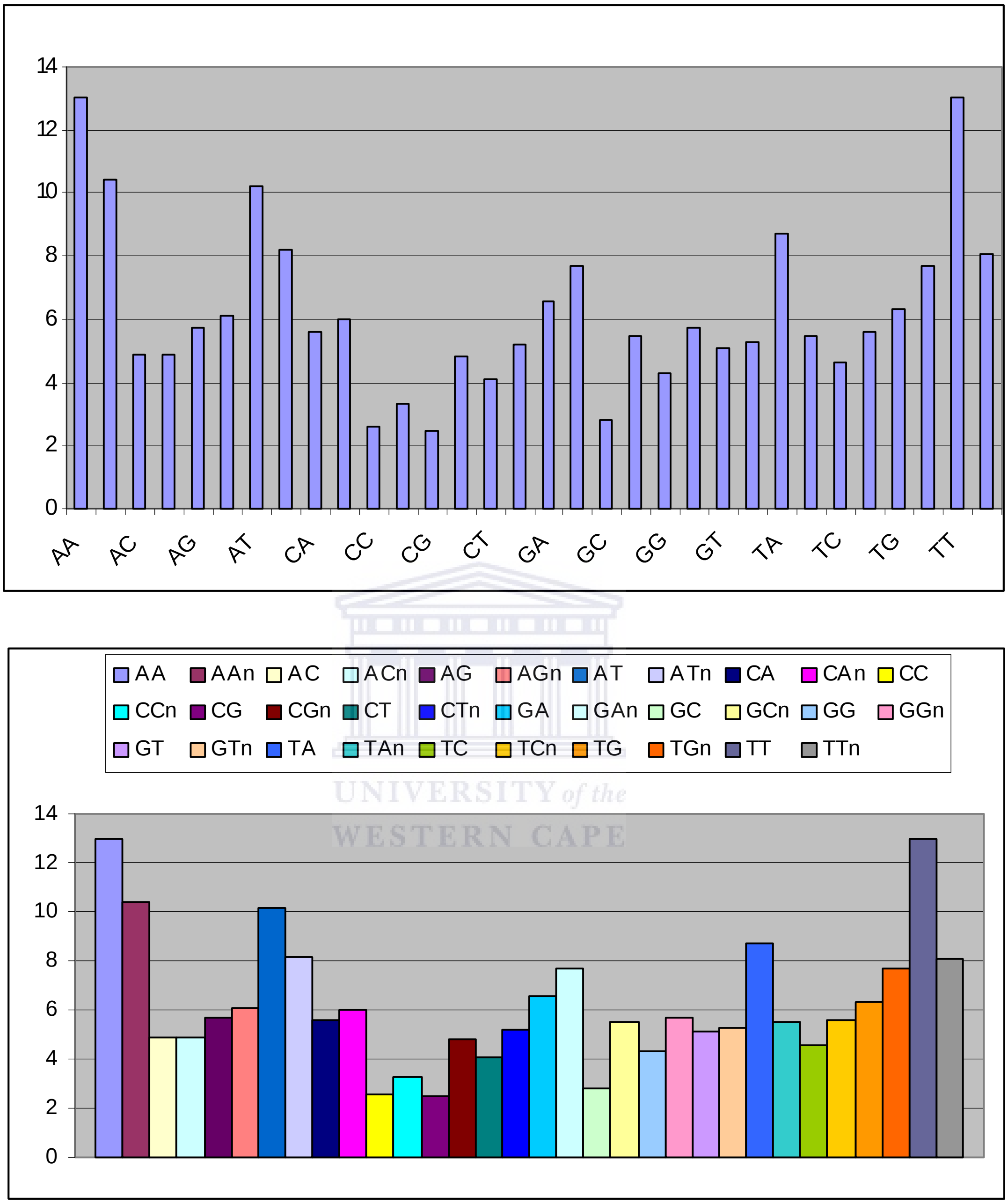
UNIVERSITY of the
WESTERN CAPE

B



UNIVERSITY of the
WESTERN CAPE

C



UNIVERSITY of the
WESTERN CAPE

Figure 5.2. Graphical representation of the dinucleotide content of promoter and non-promoter data of *E.coli* (A*) B.subtilis* (B) *and Mycobacterium* (C). Dinucleotides with the letter ‘n’ (e.g. ATn) represent dinucleotides from non-promoter sequences of the respective organisms. The same information is represented in two different graphs. The graphs depict similar dinucleotide sets (side by side) from promoter and non-promoter sets respectively.

UNIVERSITY of the
WESTERN CAPE

In *E.coli* promoters, AA, TT and AT are the predominant dinucleotide pairs (9.3%, 10.8 % and 8% respectively). The remaining dinucleotides have percentage composition around 6% (±0.5%) with GG having the lowest representation at 3.7% (figure 5.2A). *E.coli* coding sequence (used as a non-promoter in the comparison) reveals elevated CG (7.9%) and GC (8.4%) in the sequence dataset. The following dinucleotides, AA (13%), AT (10.2%) and TT (13%) are also relatively higher in *B.subtilis* promoter dataset. The proportion of AT (8%) and TA (10%) are also relatively high compared to the other dinucleotides. Percentage compositions of most dinucleotides in the promoter set are well below 6% with CC as low as 2.4%. Distribution in *B.subtilis* non-promoter sequences is more uniform compared to that of the promoter dataset, even though, dinucleotides AA and TT stand out amongst the rest of the dinucleotides. Again, the relative abundance of the two dinucleotides (AA and TT) is not comparable to their equivalents in the promoter data. There is a relative sharp rise in CC (3.3%) numbers in the non-promoter (coding sequence) data as compared to percentage in the promoter data (2.6%).

The dinucleotides CG, GC, GG, CC and TC are more prevalent in both *Mycobacterium* promoter and non-promoter sequences. The percentage compositions of these dinucleotides (CG, GC, GG and CC) are higher in the non-promoter data. A smaller proportion of the AT-rich dinucleotides TA, TT, AA and AT are usually more predominant in promoters and are even lower in the non-promoter datasets. The dinucleotide composition in the mycobacterium promoter datasets is more of a reflection of the composition in the non-promoter promoter data with less numbers of dinucleotides resulting from *W* (A and/or T). A critical analysis of the graph in figure 5.2 reveals a non-uniform distribution of dinucleotides in all the three promoter/non-promoter datasets.

UNIVERSITY of the
WESTERN CAPE

Percentage GC content analysis was performed on the entire genomes (Genbank version 111) of the three organisms i.e. *E.coli*, *B.subtilis and M.tuberculosis*. The percentage GC composition was 50%, 44% ~66% for *E.coli, B.subtilis and M.tuberculosis* respectively. The analysis of the promoter data and non-promoter data used in the study revealed a different GC content in both promoters and non-promoters. The percentage GC content of the promoter/non-promoter datasets were 43%/52% for *E.coli*, 33%/45% for *B.subtilis* and 58%/65% for *M.tuberculosis*. The figures from the analysis emphasize the point made earlier concerning the distribution of nucleotides in coding and non-coding sections of genomes. Clearly, a system that apportions some score/values (higher for certain particular dinucleotides e.g. prevalent in promoters) would seem to be an effective way of detecting promoter data from non-promoter data. One approach would be to award higher points to dinucleotides prevalent in promoters and associating higher score of a test sequence with a hypothetical promoter.

5.4.3. DINUCLEOTIDE FREQUENCY DISTRIBUTION ANALYSIS (DFDA).

The percentage composition of the dinucleotides of each organism’s promoter and non-promoter dataset was used to generate a hash table of dinucleotides for the particular organism. In order to develop a system of measure to predict promoter sequences from non-promoter sequences, the percentage composition of each dinucleotide in the non-promoter dataset was subtracted from the corresponding percentage in the promoter dataset.

UNIVERSITY of the
WESTERN CAPE

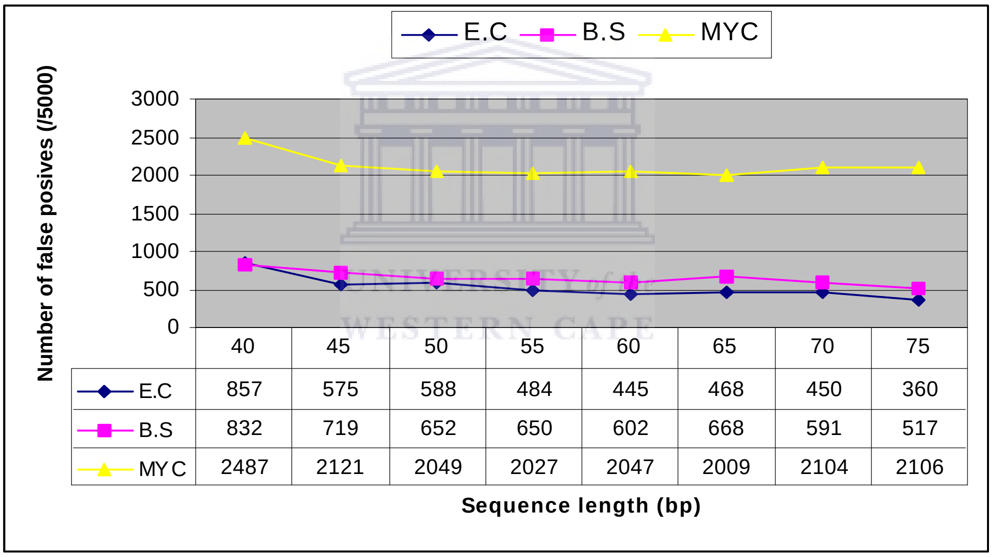
$$D_tv = D_tp - D_tnp. \qquad 5.2$$

Where D_tv is the dinucleotide value of the D_t in the hash table; D_tp and D_tnp represent the percentage composition of the dinucleotide D_t in promoter and non-promoter sequences respectively. A test sequence is analyzed by adding up the respective hash table values of all the dinucleotides in the sequence. A threshold selected by applying the measure on actual promoter sequences can then be used as a cut-off in the prediction. The dinucleotide score for the sequence in question will

$$S_c = \sum_{i=1}^{n} D_i v$$

be:

Where S_c is the aggregate of the hash table values of all the dinucleotides found in the test sequence as a 2-bp window is moved a 1 bp to the end.



	40	45	50	55	60	65	70	75
E.C	857	575	588	484	445	468	450	360
B.S	832	719	652	650	602	668	591	517
MYC	2487	2121	2049	2027	2047	2009	2104	2106

UNIVERSITY of the
WESTERN CAPE

Fig.5.3. Results indicating the number of false positives obtained from using the differences in dinucleotide content of promoter non-promoter datasets of *E.coli (Ec), B.subtilis (Bs) and Mycobacterium* respectively. Five thousand (5000) non-promoter sequences of 101 bp were used in the test set for each of the three organisms. Threshold values that resulted in 90% True Positive (using respective known promoter sequences for each organism were used to categorize test

sequences as predicted promoter sequences or non-promoter sequences. The actual data is found at the bottom of graph.

Figure 5.3 illustrates the results obtained by using such analysis. The false positive scores for *E.coli* and *B.subtilis* are quite impressive. The best scores (least number of false positives) are 360/5000 and 517/5000 for *E.coli* and *B.subtilis* respectively. These scores are fact comparable to those obtained for such recognized prediction system as neural network. Like the other results obtained on neural network and HMM systems, the results obtained for *Mycobacterium* are not quite as good. As discussed earlier in the previous two chapters, this might be attributed to a number of reasons including minimal test promoter dataset and using estimated (refer to section on *M.tuberculosis* promoter testdata) region based on either known –10 or -35 for true positive tests (promoter test data annotated not conclusive). Though best prediction score obtained on *M.tuberculosis* is 2009/5000 (40%), the score is still less than 50% and therefore DFDA has the potential of being used as a promoter prediction tool.

UNIVERSITY of the
WESTERN CAPE

5.4.4. TRIPLET COMPOSITION ANALYSIS

The nucleotide composition analysis has been extended to triplet from dinucleotide as shown in table 5.3 and figure 5.3. Subtle but detectable differences between the composition of most of the triplets in the different data are observed in all three bacteria. Some triplets including ATT, ATA, CAT, TAT, TTT, TTA, CGC, CTG, GCG and GCG were found to vary considerably in composition (+1%) between the promoter sequences and coding (non-promoter) sequences.

An examination of the above table (table 5.3) reveals triplet with higher proportion of A or T (*W*) to be relatively higher in composition in promoter data as compared to non-promoter data. A few of such triplets are highlighted in bold (table 5.3). The higher composition of *W* nucleotides for promoter data is even true for the high GC-rich mycobacterium data. On the other hand, GC-rich triplets tend to be more predominant in non-promoter sequences compared to promoter sequences in each of the three organisms, figure 5.3. The differences in composition of some of the AT dominated and GC dominated nucleotides in promoter/non-promoter data are in some cases very significant. Examples include AAA, ATA, ATT, TAT, TTA, TTT in *E.coli* and *B.subtilis* and CGC, TCG among others in Mycobacterium. The pattern of distribution of triplets in *B.subtilis* promoter and

	E.coli		***B.subtilis***		***Mycobacterium***	
Triplet	**P**	**NP**	**P**	**NP**	**P**	**NP**
AAA	**5.5**	**3.9**	**3.8**	**1.6**	0.9	0.7
AAC	1.6	1.8	1.8	2.5	1.4	1.6
AAG	2.2	2.5	1.8	1.6	1.5	1.2
AAT	3.6	2.2	1.9	1.4	0.5	0.6
ACA	2.4	1.7	2.0	1.3	1.3	1.3
ACC	0.7	0.9	0.7	1.6	2.0	2.3
ACG	0.7	1.2	1.1	2.0	1.6	2.2
ACT	1.1	1.1	1.0	1.2	1.4	1.1
AGA	1.9	1.9	1.9	1.1	1.1	0.8
AGC	0.9	1.7	2.0	1.4	1.7	1.4
AGG	1.7	1.4	1.3	0.8	1.5	1.1
AGT	1.3	1.1	1.2	0.9	1.3	0.6
ATA	**3.5**	**1.5**	1.7	1.0	1.1	0.3
ATC	1.3	2.0	1.7	1.4	1.4	1.9
ATG	2.0	2.4	1.7	2.2	1.2	1.2
ATT	**3.4**	**2.3**	**3.1**	**1.9**	0.8	0.7
CAA	2.0	2.2	1.6	1.3	1.3	2.0
CAC	1.0	0.9	1.7	1.4	1.5	2.3
CAG	0.9	1.4	1.8	1.3	1.9	1.3
CAT	1.8	1.6	**2.1**	**1.1**	1.1	1.3
CCA	0.7	0.9	0.5	1.2	1.6	1.7
CCC	0.5	0.3	0.9	0.3	2.0	1.6
CCG	0.5	1.1	0.8	2.1	3.3	3.5
CCT	0.9	0.9	1.2	1.1	1.1	1.2
CGA	0.8	1.1	1.1	1.9	2.6	3.1
CGC	0.4	1.0	**0.9**	**2.9**	**2.4**	**3.6**
CGG	0.6	1.5	0.7	2.3	3.4	3.5
CGT	0.7	1.2	1.4	1.8	1.5	1.6
CTA	0.9	0.9	1.6	0.8	0.9	0.6
CTC	0.7	0.7	1.1	0.8	1.0	1.5
CTG	**0.7**	**1.9**	1.9	2.4	1.9	1.9
CTT	1.8	1.8	1.2	1.3	0.8	1.0

UNIVERSITY of the
WESTERN CAPE

GAA	2.4	2.8	**1.6**	**2.7**	1.6	1.2
GAC	0.9	1.0	0.6	1.1	2.4	2.3
GAG	1.5	1.5	1.7	1.0	1.4	1.2
GAT	1.7	2.3	1.9	1.7	2.0	1.5
GCA	0.8	1.6	2.4	1.4	1.4	2.4
GCC	0.5	1.1	0.6	2.3	2.7	2.7
GCG	0.6	1.2	**1.5**	**3.5**	2.7	3.2
GCT	0.9	1.6	1.7	1.7	1.3	1.9
GGA	1.7	1.9	1.0	1.4	1.9	1.2
GGC	0.5	1.4	**1.3**	**2.6**	2.7	3.4
GGG	0.9	1.0	0.6	1.4	2.1	1.5
GGT	1.1	1.4	1.0	1.8	2.4	2.2
GTA	1.4	1.3	0.8	1.2	0.9	0.6
GTC	0.7	1.1	1.0	1.0	2.0	2.0
GTG	1.2	1.3	0.7	1.7	2.1	1.8
GTT	1.9	1.7	2.5	1.3	1.6	1.3
TAA	3.1	1.5	2.5	1.2	0.7	0.2
TAC	1.3	1.2	0.8	1.1	1.0	0.6
TAG	1.2	0.7	1.0	0.3	0.8	0.2
TAT	**3.1**	**2.0**	2.3	1.8	0.9	0.7
TCA	1.7	1.9	2.3	1.3	1.5	1.5
TCC	0.9	1.0	1.1	0.5	1.3	1.4
TCG	0.7	1.2	0.8	1.3	**2.1**	**3.1**
TCT	1.3	1.5	1.9	1.3	0.8	0.9
TGA	2.2	2.8	1.9	2.1	1.8	1.1
TGC	0.9	1.4	1.8	2.1	1.6	1.8
TGG	1.1	1.8	**1.4**	**2.8**	2.0	2.1
TGT	2.1	1.7	1.3	0.8	1.3	1.2
TTA	**3.0**	**1.9**	2.4	1.6	0.5	0.2
TTC	1.8	1.8	2.3	1.2	1.3	1.6
TTG	2.4	2.2	2.0	1.6	1.5	1.3
TTT	**5.9**	**2.3**	3.1	2.3	0.9	0.7

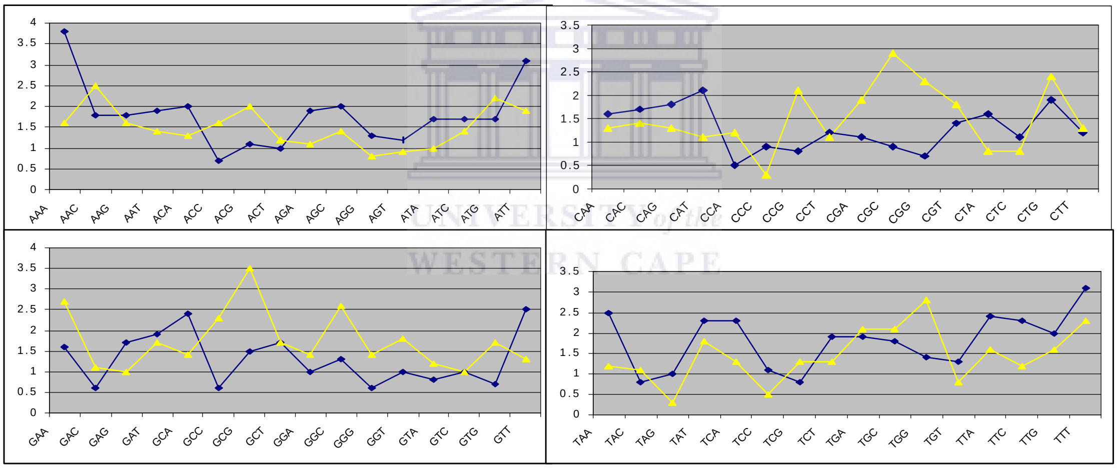
UNIVERSITY of the
WESTERN CAPE

Table 5.3. Percentage composition of all sixty-four triplets in promoter (P) and non-promoter (NP) of the three organisms namely *E.coli*, *B.subtilis* and *Mycobacterium*. Equal sizes (numbers and fragment sizes) of nucleotides as arranged in their respective genome were analyzed. Triplets with difference of one percent or more (+1%) are highlighted in bold.

non-promoter appears to be similar to those of *E.coli*. Since both have high AT-rich chromosomes.

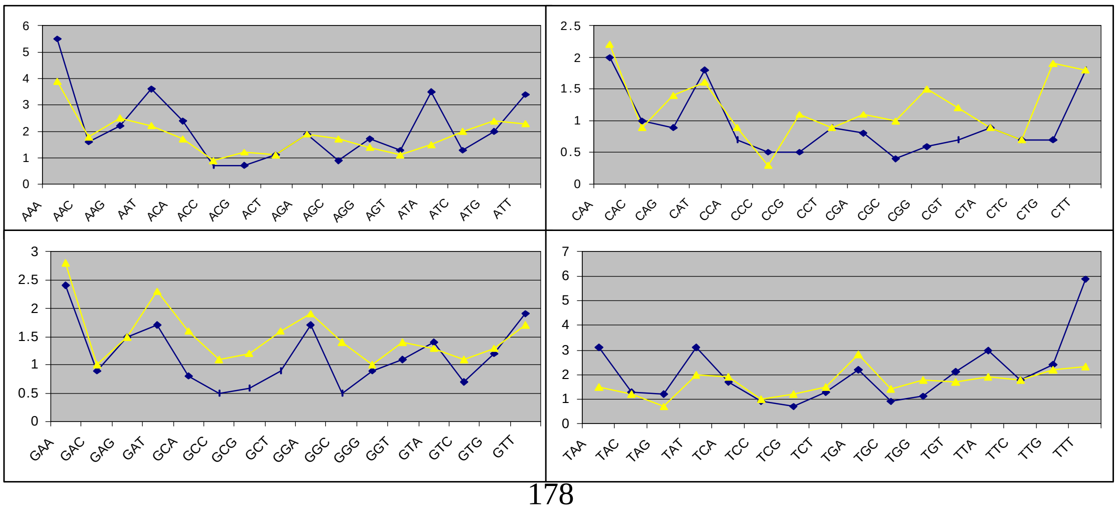
However, the actual composition/content of triplets vary in frequency. Certain triplets in both sets of data i.e. promoter and non-promoter appear to be insignificant in almost all the organisms with respect to its composition. Such triplets include ACG, GTA, CCT and CTT; may play different role(s) that may have nothing to do with the quantity in the promoter region. Triplet analysis, just like the dinucleotide sequence analysis revealed a clear difference in nucleotide composition between promoters and non-promoters in the respective organisms. These distinctive differences may be utilized to develop a system capable of distinguishing promoter sequences of an organism from its coding sequences, as was the case with the dinucleotides.

A



UNIVERSITY of the
WESTERN CAPE

B



C

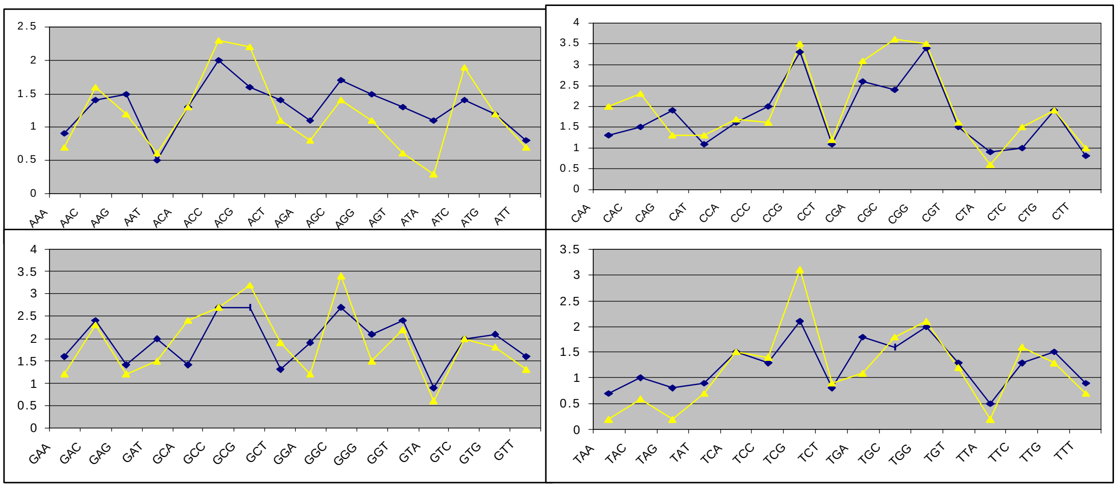


Figure 5.4. Distribution (percentage composition) of the sixty-four (64) possible triplets in *E.coli* promoter (square/blue plot) and non-promoter (triangle/yellow) data set (A), *B.subtilis* data set (B) and Mycobacteria data set (C). Variations in the distribution of certain types of triplets are evident in the two data sets of promoter/non-promoter. Triplets that are relatively prevalent in both data include AAA, ATT and TTT whereas the triplets GCG, GCC and CGG fluctuate widely in composition between the two sets of data. Other triplets ACT, CCT, CTT and GTA are consistently found to have almost the same composition in all data sets in the three organisms.

UNIVERSITY of the
WESTERN CAPE

5.4.5. Triplet Frequency Distribution Analysis on *E.coli*.

UNIVERSITY of the WESTERN CAPE

Results from prediction using hash table generated from triplet frequency distribution of same fragment size as test data, fixed 75-bp fragment sizes and fixed 101 bp fragment sizes are shown in table 5.5, 5.6 and 5.7 respectively. Corresponding graphs are shown in figures 5.5A, 5.5B and 5.5C. The trend is similar to those observed for neural network and HMM prediction. Once again, there is no obvious direct correlation between fragment size and predictability. However, the overall results are better (less number of false positives) than results obtained from both HMM and Neural network. Triplet Frequency Distribution analysis is favored because it gives more variables for the hash table (64) as compared to sixteen (16) for dinucleotide frequency analysis.

SETS	1	2	3	4	5	Av.	%
10_40	1009	994	968	963	994	985.6	19.7
10_45	646	703	655	678	657	667.8	13.4
10_50	642	625	653	620	620	632.0	12.6
10_55	656	627	635	651	626	639.0	12.8
10_60	462	473	485	492	453	473.0	9.5
10_65	488	485	500	494	497	492.8	9.9
10_70	564	553	565	557	569	561.6	11.2
10_75	490	474	504	491	498	491.4	9.8
20_40	971	939	936	948	943	947.4	18.9
20_45	543	550	535	559	541	545.6	10.9
20_50	588	585	588	570	561	578.4	11.6
20_55	535	528	520	526	515	524.8	10.5
20_60	514	516	503	524	504	512.2	10.2
20_65	382	373	379	385	382	380.2	7.6
20_70	469	444	444	447	433	447.4	8.9
20_75	345	353	361	344	353	351.2	7.0
30_40	767	734	747	753	770	754.2	15.1
30_45	637	655	622	650	638	640.4	12.8
30_50	580	578	567	563	558	569.2	11.4
30_55	436	425	426	420	431	427.6	8.6
30_60	311	330	313	312	290	311.2	6.2
30_65	400	398	405	409	408	404.0	8.1
30_70	353	343	355	355	354	352.0	7.0
30_75	300	312	333	324	335	320.8	6.4
40_40	900	876	844	857	884	872.2	17.4
40_45	533	524	500	515	514	517.2	10.3
40_50	543	568	571	546	558	557.2	11.1
40_55	462	463	444	454	459	456.4	9.1

UNIVERSITY of the
WESTERN CAPE

40_60	381	393	396	383	351	380.8	7.6
40_65	388	388	386	374	376	382.4	7.6
40_70	334	307	315	312	309	315.4	6.3
40_75	349	349	368	344	335	349.0	7.0
50_40	712	701	677	671	690	690.2	13.8
50_45	584	592	582	593	583	586.8	11.7
50_50	496	522	494	508	495	503.0	10.1
50_55	448	447	413	441	433	436.4	8.7
50_60	444	451	429	455	429	441.6	8.8
50_65	392	384	385	377	382	384.0	7.7
50_70	407	371	371	375	376	380.0	7.6
50_75	356	356	371	358	362	360.6	7.2

Table 5.4. False positive results obtained from the individual hash tables generated from promoter and non-promoter sequences of the same size (number of sequences and sequence lengths). Tested sequences have the same fragment sizes as the sets (promoter/non-promoter) used to develop the table. Five random sequences were generated from each of the original test sequences (101 bp) to obtain results very reflective on the actual test data.

UNIVERSITY of the
WESTERN CAPE

SETS	1	2	3	4	5	Av.	%
10_40	595	592	618	600	605	602.0	12.0
10_45	586	593	604	586	610	595.8	12.0
10_50	595	600	615	592	591	598.6	12.0
10_55	555	549	563	545	568	556.0	11.0
10_60	452	442	450	450	459	450.6	9.0
10_65	430	432	439	437	450	437.6	8.8
10_70	544	544	564	557	559	553.6	11.1
10_75	476	458	483	469	480	473.2	9.5
20_40	446	451	452	443	449	448.2	9.0
20_45	418	417	428	414	427	420.8	8.4
20_50	357	351	350	341	347	349.2	7.0
20_55	440	446	460	450	449	449.0	9.0
20_60	500	493	512	504	487	499.2	10.0
20_65	392	409	402	398	409	402.0	8.0
20_70	344	358	359	351	338	350.0	7.0
20_75	370	380	381	369	372	374.4	7.5
30_40	491	507	509	502	521	506.0	10.1
30_45	481	500	513	490	510	498.8	10.0
30_50	446	453	455	443	463	452.0	9.0
30_55	354	346	359	355	352	353.2	7.1
30_60	292	284	302	304	295	295.4	5.9
30_65	346	337	351	352	355	348.2	7.0
30_70	324	318	332	329	333	327.2	6.5
30_75	324	331	353	345	355	341.6	6.8
40_40	426	412	421	404	408	414.2	8.3
40_45	356	361	371	359	356	360.6	7.2
40_50	408	397	408	382	398	398.6	8.0
40_55	420	418	420	409	408	415.0	8.3
40_60	352	345	366	341	332	347.2	6.9
40_65	350	360	362	357	335	352.8	7.1
40_70	319	318	334	321	305	319.4	6.4

40_75	333	327	349	330	306	329.0	6.6
50_40	312	311	329	311	322	317.0	6.3
50_45	347	349	373	360	358	357.4	7.1
50_50	314	303	321	311	315	312.8	6.3
50_55	334	347	359	343	349	346.4	6.9
50_60	405	417	430	415	439	421.2	8.4
50_65	391	412	406	408	418	407.0	8.1
50_70	357	373	370	357	354	362.2	7.2
50_75	313	322	329	316	317	319.4	6.4

Table 5.5. The procedure used to obtain the data is similar to that used to obtain results in table 5.4. However, datasets have sequences of 75 bp fragment size each. The average numbers of false positives together with their respective percentage are shown in columns seven and eight.

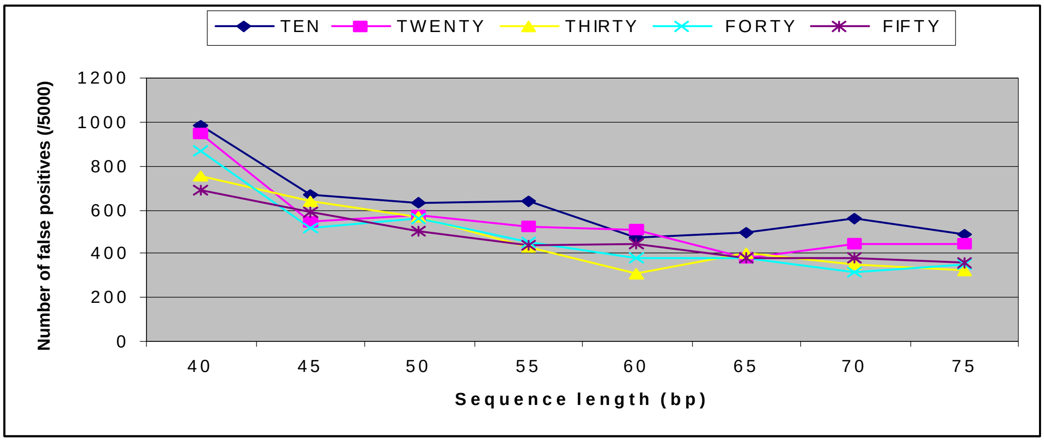
	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	701	521	329	451	418
45	548	496	446	397	382
50	650	407	290	455	334
55	680	518	276	326	301
60	496	518	229	304	357
65	470	352	321	310	488
70	502	344	271	274	366
75	315	400	281	317	336

UNIVERSITY of the
WESTERN CAPE

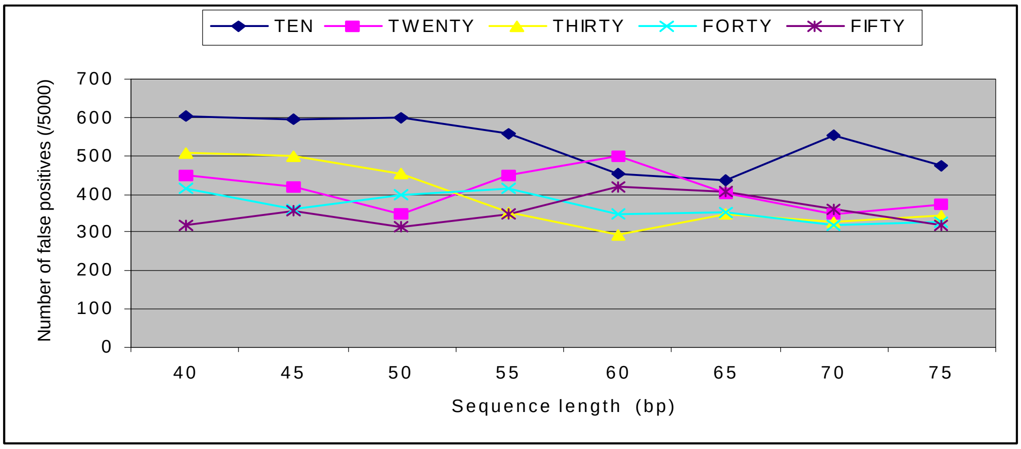
Table 5.6. Triplet frequency distribution analysis results on five thousand E.coli non-promoter data of 101 bp fragment size. A cut-off value that resulted in 90% TP (true positive) was manually selected and used as prediction threshold.

UNIVERSITY of the
WESTERN CAPE

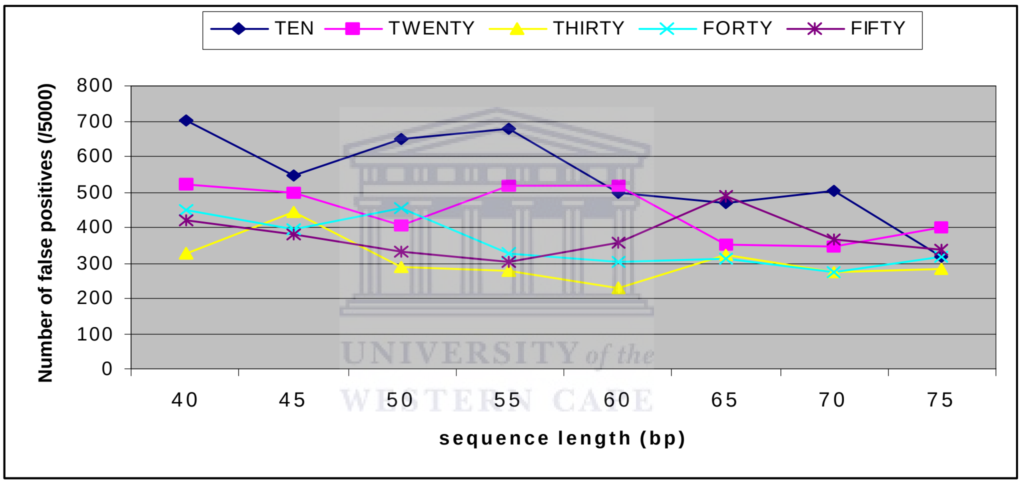
A



B



C



UNIVERSITY of the WESTERN CAPE

Figure 5.5. Graphs of results shown in table 5.3 (A), 5.4 (B) and 5.5 (C) which represent the number of false positives obtained by using hash table values from designed sequence sets on sequences of the same fragment size (A), of 75 bp fragment size (B) and 101 bp fragment sizes (C). In all instances, cut-off values that represented 90% true positive were used to determine which test sequences were considered predicted promoter sequences.

UNIVERSITY of the
WESTERN CAPE

A consistent trend is observed in the first type of designed set where the test sequences are varied depending on which sequence subset was used to generate the hash table. An increase in size of test sequences from 40 bp results in decrease in number of false positives. Also, the differences in the results do not fluctuate as seen with the previous methods of neural network and hidden Markov model. Percentage false positive results ranges from a high 19.7% (10_40) to an impressive low of 6.2% for hash table values developed from a sequence subset of thirty sequences of 60 bp (30_60) fragment size each. As in NN and HMM methods, results were better on test sequences fixed at 75 bp, which emphasizes a belief from this study that, a longer region of 75 bp is probably the ideal practical fragment size that should be used in promoter prediction/detection. A worst result of 12% false positives is reflected on sub sequence set 10_40 with indistinguishable results from 10_45 . However, the best result, also derived from 30_60 (thirty sequences of 60 bp fragment sizes each) of 5.9% is even better than that obtained for the previous test set of 60 bp. A consistent trend in all the three methods of prediction has been the overall better results as test sequences are increased from a fixed 75 bp to 101.

.

5.4.6. Triplet Frequency Distribution analysis on *B.subtilis*

The same triplet frequency analysis procedure used on *E.coli* data was applied to *B.subtilis*. The sequence subsets used are the same as those used for HMM and NN analysis. Results from this analysis is very similar to those of *E.coli* with hash values from sequence subset of fifty (50) producing the best results (least number of false positives) as compared to thirty (30) in *E.coli*. False positives range from slightly fewer than eight hundred (800) to just above four hundred (400), except for the false positive results obtained from 10_60 (1003).

UNIVERSITY of the
WESTERN CAPE

	1	2	3	4	5	Av.	%
10_40	673	629	614	633	647	639.2	12.8
10_45	833	746	745	728	741	758.6	15.2
10_50	891	726	737	719	703	755.2	15.1
10_55	603	495	504	515	510	525.4	10.5
10_60	1131	1004	979	1003	992	1021.8	20.4
10_65	748	576	560	555	550	597.8	12.0
10_70	664	538	527	537	520	557.2	11.1
10_75	644	497	484	477	475	515.4	10.3

20_40	715	620	612	588	612	629.4	12.6
20_45	592	515	499	492	487	517.0	10.3
20_50	931	784	793	797	799	820.8	16.4
20_55	651	525	537	541	544	559.6	11.2
20_60	796	655	655	653	661	684.0	13.7
20_65	817	691	668	666	690	706.4	14.1
20_70	857	676	689	694	679	719.0	14.4
20_75	694	539	530	538	543	568.8	11.4
30_40	787	679	678	683	703	706.0	14.1
30_45	688	583	538	578	547	586.8	11.7
30_50	684	586	575	562	557	592.8	11.9
30_55	672	512	540	535	531	558.0	11.2
30_60	757	636	604	612	592	640.2	12.8
30_65	684	535	541	526	547	566.6	11.3
30_70	743	551	563	578	566	600.2	12.0
30_75	852	662	679	674	665	706.4	14.1
40_40	615	557	567	535	553	565.4	11.3
40_45	554	492	469	462	445	484.4	9.7
40_50	715	599	588	563	585	610.0	12.2
40_55	629	527	545	557	542	560.0	11.2
40_60	668	566	562	559	570	585.0	11.7
40_65	855	740	709	722	728	750.8	15.0
40_70	725	593	590	603	585	619.2	12.4
40_75	715	551	560	559	559	588.8	11.8
50_40	634	606	569	573	594	595.2	11.9
50_45	737	629	608	614	621	641.8	12.8
50_50	675	562	524	542	538	568.2	11.4
50_55	630	503	501	524	511	533.8	10.7
50_60	709	613	575	580	583	612.0	12.2
50_65	642	545	547	542	536	562.4	11.2
50_70	715	565	576	586	589	606.2	12.1
50_75	572	415	434	424	420	453.0	9.1

UNIVERSITY of the
WESTERN CAPE

Table 5.7. False positive results obtained five thousand (5000) non-promoter sequences using triplet frequency analysis. All the test sequences used had same fragment sizes as those used to generate their respective triplet hash values. Threshold values that resulted in 90% true positive for the 83 actual promoters used were used to judge the respective test sequences.

	1	2	3	4	5	Av.	%
10_40	616	616	525	511	520	557.6	11.2
10_45	1012	1012	906	885	894	941.8	18.8
10_50	834	834	651	649	649	723.4	14.5
10_55	731	731	587	558	570	635.4	12.7
10_60	1189	1189	1058	1024	1041	1100	22.0
10_65	746	746	568	565	562	637.4	12.7
10_70	856	856	673	678	666	745.8	14.9
10_75	644	644	484	477	475	544.8	10.9

20_40	876	876	691	701	691	767.0	15.3
20_45	813	813	651	652	642	714.2	14.3
20_50	842	842	654	689	681	741.6	14.8
20_55	813	813	655	661	651	718.6	14.4
20_60	703	703	556	552	548	612.4	12.2
20_65	979	979	774	778	781	858.2	17.2
20_70	831	831	641	645	649	719.4	14.4
20_75	694	694	530	538	543	599.8	12.0
30_40	614	614	488	484	477	535.4	10.7
30_45	668	668	499	506	489	566.0	11.3
30_50	530	530	409	411	391	454.2	9.1
30_55	671	671	502	505	505	570.8	11.4
30_60	747	747	576	572	577	643.8	12.9
30_65	711	711	531	534	546	606.6	12.1
30_70	684	684	522	525	514	585.8	11.7
30_75	852	852	679	674	665	744.4	14.9
40_40	694	694	545	543	525	600.2	12.0
40_45	595	595	471	460	463	516.8	10.3
40_50	806	806	649	653	646	712.0	14.2
40_55	731	731	594	589	589	646.8	12.9
40_60	663	663	526	541	531	584.8	11.7
40_65	731	731	599	597	592	650.0	13.0
40_70	641	641	504	511	499	559.2	11.2
40_75	715	715	560	559	559	621.6	12.4
50_40	555	555	452	450	441	490.6	9.8
50_45	603	603	487	466	461	524.0	10.5
50_50	531	531	416	411	415	460.8	9.2
50_55	607	607	476	473	470	526.6	10.5
50_60	538	538	441	437	433	477.4	9.5
50_65	606	606	491	490	475	533.6	10.7
50_70	667	667	538	542	538	590.4	11.8
50_75	572	572	434	424	420	484.4	9.7

UNIVERSITY of the
WESTERN CAPE

Table 5.8. False positives resulting from using generated hash tables from the various sequence subsets. Each test sequence had a sequence length of 75 bp. Five random sequences were generated from every test sequence. The average is then used to represent the number of false positives.

	TEN	TWENTY	THIRTY	FORTY	FIFTY
40	558	568	458	358	385
45	770	421	499	507	364
50	725	595	414	483	362
55	534	401	337	503	293
60	843	467	465	392	357
65	506	512	405	492	368
70	437	513	418	397	322
75	432	463	360	398	351

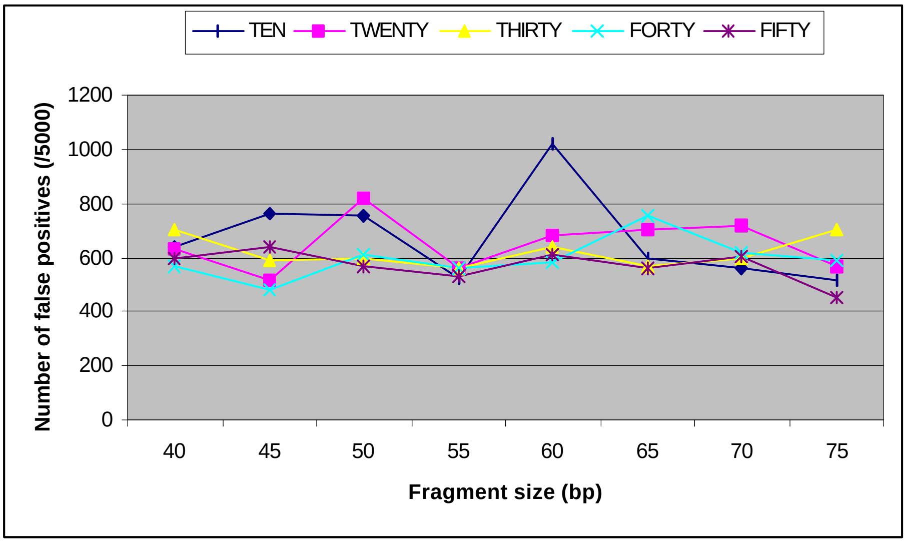
Table 5.9. Sequence length of test data sets used is 101 bp each. Total number of test sequences is 5000.

UNIVERSITY of the
WESTERN CAPE

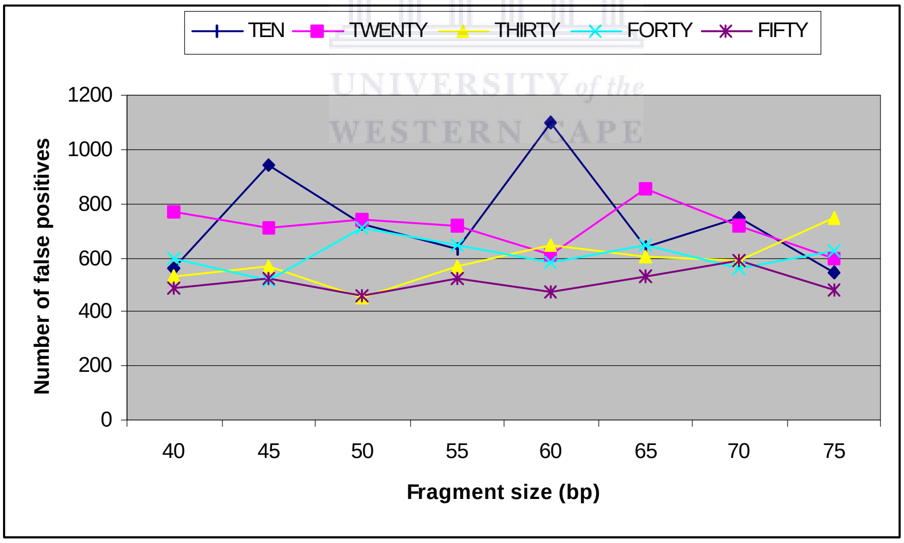
Results obtained by testing sequences of the same sequence length, 75-bp sequence length, and 101-bp sequence length, figure 5.6A, 5.6B and 5.6C respectively are not very different from that corresponding to *E.coli*. As expected, the worst results were obtained from subsets made up of only ten sequences. However, unlike the case of *E.coli*, the best results in all the three categories of test came from the set of fifty sequences. The overall results from test sequences of fragment size 101 were better than test sequences of fragment size 75 bp which were also better than test sequences of the same length as subsets used to generate their respective hash table of scores.

UNIVERSITY of the
WESTERN CAPE

A

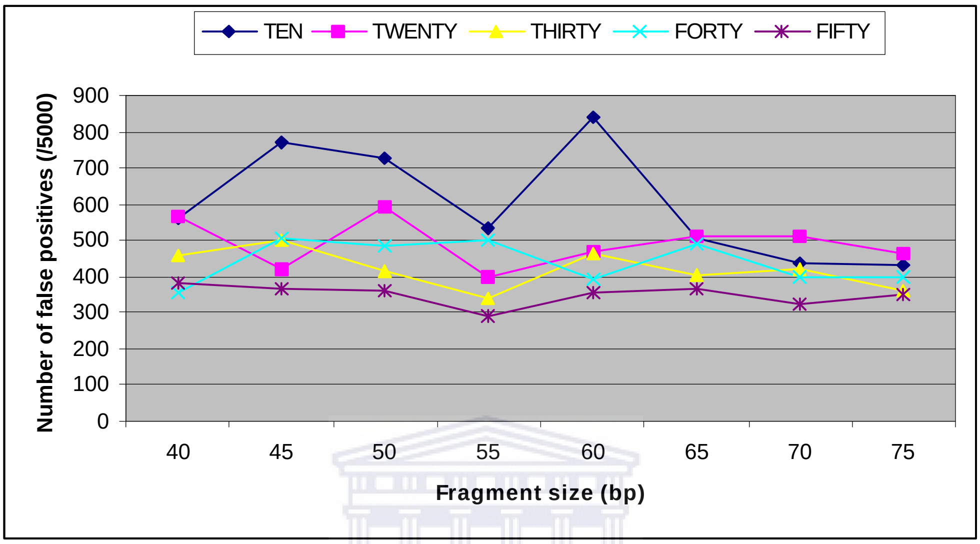


B



UNIVERSITY of the
WESTERN CAPE

C



UNIVERSITY of the
WESTERN CAPE

Figure 5.6. Graphs of results shown in table 5.8 (A), 5.9 (B) and 5.10 (C) . The three graphs represent the number of false positives obtained by using hash table values from designed sequence sets on sequences of the same fragment size (A), of 75 bp fragment size (B) and 101 bp fragment sizes (C). In all instances, cut-off values that represented 90% true positive were used to determine which test sequences were considered predicted promoter sequences. Five thousand (5000) *B.subtilis* test promoter sequences were used.

5.4.7. Triplet Frequency Analysis on Mycobacterium data (Promoter and Non-promoter).

Results are less encouraging and the explanation/reasons have been mentioned in the previous discussions on HMM and NN methods on mycobacterium. The trend in the three types of test as depicted in figure 5.7A, 5.7B and 5.7C is the same as observed in the other two methods of test with test sequences of 101 bp producing the best results. Sequence subset 30_50 resulted in best results for the first two tests generating false positive values of 26.9 % and 31% respectively. However, test on sequence length of 101 bp produced best result from 30_60.

UNIVERSITY of the
WESTERN CAPE

Sets	**1**	**2**	**3**	**4**	**5**	**Av**	**%**
10_40	2297	2254	2257	2279	2286	1819.4	36.4
10_45	2318	2314	2305	2356	2315	1860.6	37.2
10_50	2075	2066	2115	2073	2073	1667.8	33.4
10_55	1996	1978	1967	1957	1947	1581.6	31.6
10_60	2062	2070	2064	2058	2084	1652.8	33.1
10_65	1908	1913	1933	1958	1934	1544.4	30.9
10_70	1854	1835	1816	1850	1823	1473	29.5
10_75	1735	1750	1744	1741	1731	1396	27.9
20_40	1858	1919	1864	1893	1880	1510.8	30.2
20_45	2310	2310	2317	2291	2328	1849.6	37.0
20_50	2252	2212	2263	2240	2223	1797.4	36.0
20_55	2419	2402	2395	2413	2399	1929.8	38.6
20_60	2428	2457	2466	2474	2452	1969	39.4
20_65	2046	2073	2038	2063	2027	1648	33.0
20_70	1939	1972	1981	2013	2026	1585	31.7
20_75	2224	2219	2232	2211	2233	1781.2	35.6
30_40	2260	2252	2304	2272	2332	1823.6	36.5
30_45	2499	2497	2474	2479	2520	1995.8	40.0
30_50	2486	2480	2474	2504	2480	1994.8	39.9
30_55	2238	2274	2256	2231	2258	1805.8	36.1
30_60	2533	2535	2543	2551	2565	2038.4	40.8
30_65	2350	2346	2352	2371	2333	1889.8	37.8
30_70	1986	1989	1967	1996	1989	1593.6	31.9
30_75	1883	1881	1903	1907	1915	1520.8	30.4
40_40	2207	2222	2201	2209	2201	1775.8	35.5
40_45	1664	1676	1669	1673	1674	1344.4	26.9
40_50	1936	1921	1920	1921	1968	1547.6	31.0
40_55	1731	1730	1719	1702	1732	1384.4	27.7
40_60	1833	1883	1843	1869	1852	1493.6	29.9
40_65	1818	1822	1840	1842	1808	1472.4	29.5
40_70	1953	1946	1945	1955	1931	1567.8	31.4
40_75	1826	1816	1840	1816	1823	1467.6	29.4
50_40	2218	2190	2218	2231	2233	1781.4	35.6
50_45	2457	2455	2428	2446	2445	1967.2	39.3
50_50	2342	2357	2374	2355	2347	1895.6	37.9

UNIVERSITY of the
WESTERN CAPE

50_55	2423	2466	2448	2424	2421	1962.2	39.2
50_60	2237	2239	2245	2255	2234	1805.2	36.1
50_65	2262	2281	2282	2289	2275	1832.8	36.7
50_70	2223	2213	2209	2244	2231	1787.8	35.8
50_75	2116	2110	2131	2131	2143	1707.6	34.2

Table 5.10. Results obtained on five sets of mycobacterium test sequences used to test the ability of TFD to discriminate against non-promoter (coding sequences). The test sequences had fragment sizes equivalent to those used in developing to the respective hash tables. The average number of false positives per 5000 and the percentage false positives are shown in the seventh and eight columns respectively.

Sets	**1**	**2**	**3**	**4**	**5**	**Av.**	**%**
10_40	**2449**	**2441**	**2460**	**2445**	**2449**	**2448.80**	**48.98**
10_45	**2459**	**2457**	**2473**	**2448**	**2459**	**2459.20**	**49.18**
10_50	**2411**	**2394**	**2411**	**2385**	**2411**	**2402.40**	**48.05**
10_55	**1810**	**1801**	**1826**	**1779**	**1810**	**1805.20**	**36.10**
10_60	**1876**	**1846**	**1866**	**1850**	**1876**	**1862.80**	**37.26**
10_65	**1852**	**1845**	**1865**	**1836**	**1852**	**1850.00**	**37.00**
10_70	**1792**	**1784**	**1774**	**1774**	**1792**	**1783.20**	**35.66**
10_75	**1735**	**1750**	**1744**	**1741**	**1735**	**1741.00**	**34.82**
20_40	**2284**	**2282**	**2264**	**2275**	**2284**	**2277.80**	**45.56**
20_45	**2214**	**2214**	**2204**	**2227**	**2214**	**2214.60**	**44.29**
20_50	**2272**	**2284**	**2270**	**2261**	**2272**	**2271.80**	**45.44**
20_55	**2152**	**2144**	**2167**	**2170**	**2152**	**2157.00**	**43.14**
20_60	**2213**	**2211**	**2223**	**2208**	**2213**	**2213.60**	**44.27**
20_65	**2331**	**2336**	**2358**	**2328**	**2331**	**2336.80**	**46.74**
20_70	**2354**	**2358**	**2368**	**2339**	**2354**	**2354.60**	**47.09**
20_75	**2224**	**2219**	**2232**	**2211**	**2224**	**2222.00**	**44.44**
30_40	**2497**	**2499**	**2492**	**2498**	**2497**	**2496.60**	**49.93**
30_45	**2381**	**2396**	**2396**	**2403**	**2381**	**2391.40**	**47.83**
30_50	**2219**	**2221**	**2222**	**2233**	**2219**	**2222.80**	**44.46**
30_55	**2023**	**2047**	**2058**	**2056**	**2023**	**2041.40**	**40.83**
30_60	**1966**	**1981**	**1988**	**1990**	**1966**	**1978.20**	**39.56**
30_65	**2014**	**2029**	**2023**	**2032**	**2014**	**2022.40**	**40.45**
30_70	**1767**	**1776**	**1791**	**1788**	**1767**	**1777.80**	**35.56**
30_75	**1883**	**1881**	**1903**	**1907**	**1883**	**1891.40**	**37.83**
40_40	**2141**	**2124**	**2146**	**2115**	**2141**	**2133.40**	**42.67**
40_45	**2037**	**2031**	**2030**	**2010**	**2037**	**2029.00**	**40.58**
40_50	**1896**	**1865**	**1886**	**1858**	**1896**	**1880.20**	**37.60**
40_55	**1885**	**1860**	**1884**	**1852**	**1885**	**1873.20**	**37.46**
40_60	**1791**	**1756**	**1797**	**1767**	**1791**	**1780.40**	**35.61**
40_65	**1948**	**1925**	**1926**	**1933**	**1948**	**1936.00**	**38.72**
40_70	**1580**	**1554**	**1574**	**1546**	**1580**	**1566.80**	**31.34**
40_75	**1826**	**1816**	**1840**	**1816**	**1826**	**1824.80**	**36.50**
50_40	**2451**	**2457**	**2447**	**2444**	**2451**	**2450.00**	**49.00**

UNIVERSITY of the WESTERN CAPE

50_45	2369	2354	2359	2361	2369	2362.40	47.25
50_50	2450	2439	2438	2449	2450	2445.20	48.90
50_55	2429	2415	2420	2439	2429	2426.40	48.53
50_60	2287	2273	2283	2267	2287	2279.40	45.59
50_65	2211	2204	2210	2224	2211	2212.00	44.24
50_70	2101	2106	2130	2116	2101	2110.80	42.22
50_75	2116	2110	2131	2131	2116	2120.80	42.42

Table 5.11. False positive results obtained on five thousand (5000) mycobacterium test sequences of 75 bp sequence-length each. In each case, threshold value which resulted in 90% True Positive (TP) was manually selected and used as the cut-off. Average score for each set and the percentage true positive values are in the seventh and the eighth columns respectively.

	TEN	**TWENTY**	**THIRTY**	**FORTY**	**FIFTY**
40	1406	1354	1268	1282	1402
45	1369	1447	1299	1334	1160
50	1307	1496	1274	1469	1156
55	1131	1526	1094	1188	1251
60	1135	1593	1068	1158	1335
65	1166	1566	1390	1201	1326
70	1223	1551	1281	1019	1283
75	1135	1584	1272	1023	1303

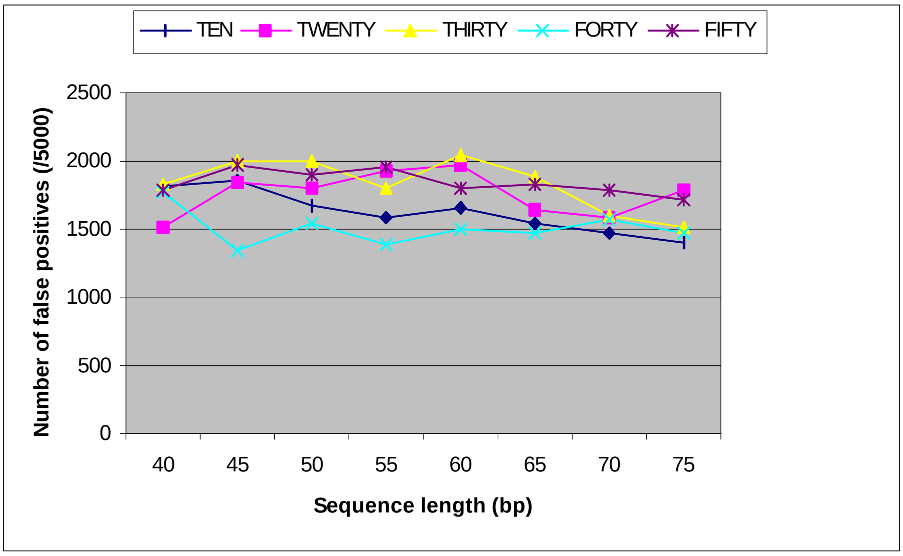
UNIVERSITY of the
WESTERN CAPE

Table 5.12. Results obtained on 5000 sets of mycobacterium test sequences using the hash models developed from the various sequence sets. Sequences tested had

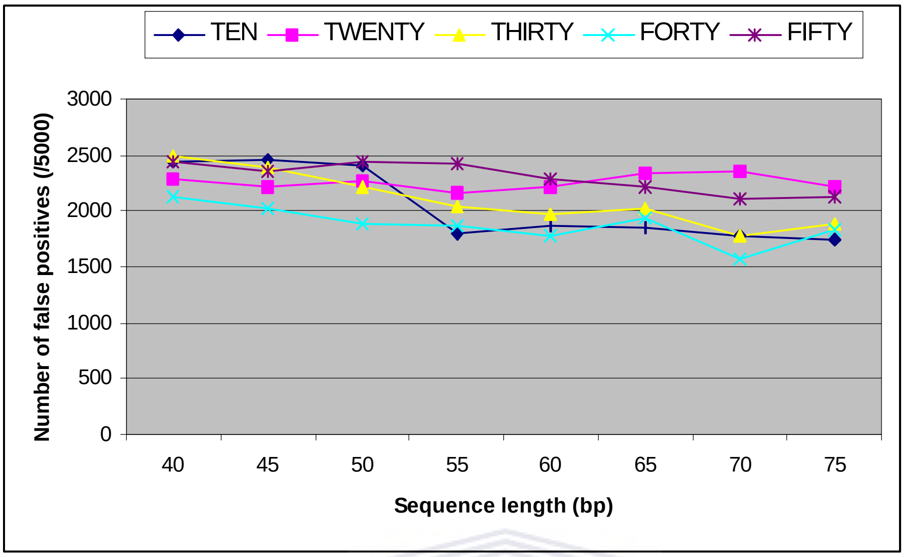
101 bp sizes. Just as in the two previous cases, a threshold was selected to obtain 90% true positive.

UNIVERSITY of the
WESTERN CAPE

A



B



UNIVERSITY of the
WESTERN CAPE

C

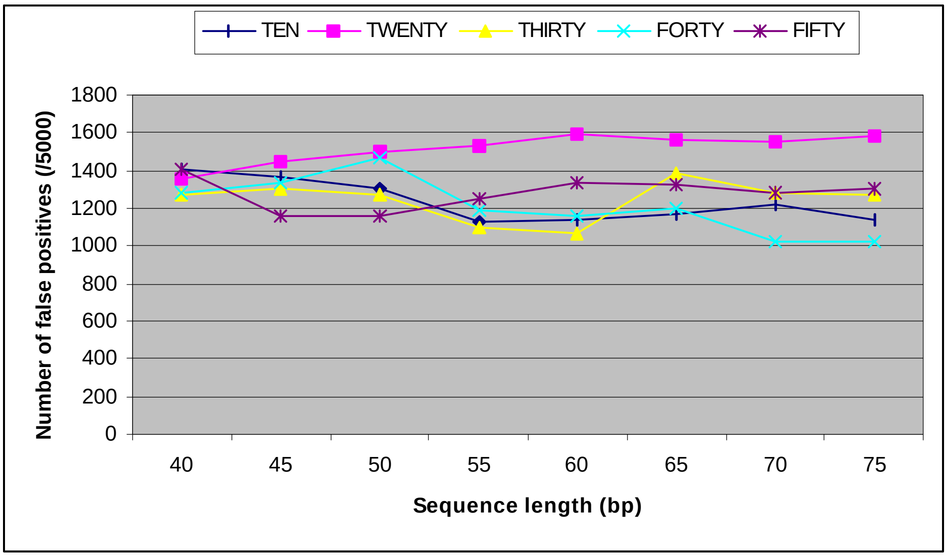


Figure 5.7. Graphs of results shown in table 5.8 (A), 5.9 (B) and 5.10 (C) . The three graphs represent the number of false positives obtained by using hash table values from designed sequence sets on sequences of the same fragment size (A), of 75 bp fragment size (B) and 101 bp fragment sizes (C). In all instances, cut-off values that represented 90% true positive were used to determine which test sequences were considered predicted promoter sequences. Five thousand (5000) *B.subtilis* test promoter sequences were used.

UNIVERSITY of the
WESTERN CAPE

Most of the knowledge gained to date on functional and regulatory elements is based on statistical analysis of experimental data. Difference in nucleotide composition of sequences is the basis of most of the computational methods used in sequence analysis. This has been exploited successfully in distinguishing promoter sequences from non-promoters. False positive prediction results of the three prokaryotes (at 90% true positives) are low enough to be used in promoter prediction of *M.tuberculosis* that has very few experimentally characterized promoters. Though false positive results on *M.tuberculosis* predictions are high, it is our opinion that, they do not really represent false positives due to the fact that the true promoter test dataset contain a lot of ‘noise’ as already discussed. Comparison of the results of DFDA to TFDA (false positive values of 360/5000, 517/5000 and 2009/5000 for *E.coli, B.subtilis* and Mycobacterium respectively to 229/5000, 293/5000 and 1068/5000 from TFDA) revealed TFDA to be a better prediction methodology compared to DFDA. The ratios of false positives above (TFDA to DFDA) are approximately 4:6, 4:6 and 3.5:6.5 for *E.coli, B.subtilis* and *M.tuberculosis* respectively. A brief statistical analysis also exposed that, there is not enough available promoter data to carry out tetranucleotide analysis, which would have resulted in 256 possible combinations instead of 64 for trinucleotides. TFDA therefore falls into the same category as ANN and HMM as very useful methodology to predict/detect promoter sequences.

WESTERN CAPE

Chapter six

Combining all three prediction systems (HMM, NN and TFDA).

ABSTRACT

The best parameters from each of the sets in the three methodologies have been combined in an attempt to optimize the predictability of the methods on promoters of *E.coli*, *B.subtilis* and most importantly, *M.tuberculosis*. Models developed on ANN, HMM and TFDA that produced best scores (least number of false positives) in the previous three chapters for *E.coli*, *B.subtilis* and *M.tuberculosis* (75 bp windows) were used. Three test datasets were constructed. The first data consisted of *E.coli* and *B.subtilis* genome sequences (80 sets each) of 481 bp with their respective eighty test promoters (101 bp each) surrounded by 190 nucleotides on either side as found in their respective genomes. Second data consisted of sections of the respective genomes of *E.coli* and *B.subtilis* (~5-10 kb), harboring three known test promoter sequences. Third data consisted of the intergenic regions of the three organisms namely, *E.coli*, *B.subtilis* and *M.tuberculosis*. Selected models were used on first and third datasets individually and then as a combined tool. Test on the first testdata using the selected models individually (not combined) resulted in 72.5%/27.5% TP/FP and 89%/11% TP/FP for *E.coli* and *B.subtilis* respectively. As combined (filtering through all three methods), 47 (59%) and 75 (82%) true positive predictions were achieved for *E.coli* and *B.subtilis* respectively. Plotting results from the test on the nucleotide regions covering ~5-10 kb of the respective genomes revealed distinct peaks at sections where the promoters were known to be located in the respective genomes of the organisms (*E.coli and B.subtilis*). Due to the nature of the results obtained using the prediction methods (individually and as combined), both types of predictions were carried out on intergenic regions in the entire genomes of *E.coli*, *B.subtilis* and *M.tuberculosis*. The results on the predictions have been made public and can be assessed at the following uniform resource locators (url): http://www.sanbi.ac.za/tb/promoters.html.

UNIVERSITY of the WESTERN CAPE

6.1. INTRODUCTION

One of the major problems frequently encountered by researchers using prediction systems is ranking and subsequent correlation of predicted results. The problem of correlation and integration of predicted results is often compounded due to the fact that, prediction systems are often based on different methods and algorithms. In the study of the prediction systems in the previous three chapters, models developed on certain sequence subsets had been expected to perform well by producing results that are consistent and good. Though models on some subsets did produce very good results, the results were not very consistent. Hence, models developed on sequence subsets that produced best results (least number of false positives) for each prediction system (figure 6.1) were selected to represent the methods. This chapter gives an insight into the attempt at developing an integrated prediction system based on the models developed on subsequences that produced best results in the previous three chapters. The resultant integrated prediction tool if successful, would be used on the entire genomes of *E.coli, B.subtilis and M.tuberculosis* in predicting promoter sequences upstream of their respective genes.

UNIVERSITY of the
WESTERN CAPE

6.2. Methods

6.2.1. Defining Promoter Prediction Region

A problem consistently encountered in this study was defining a section that constitutes promoter region. In all the three prediction systems, the average results for 75 bp fragment windows were in most cases better than tests carried on sequences of the same size as models. A 75 bp window was selected as the promoter test window in all predictions done in this chapter. With the test cases where known promoters were placed between protein coding sequences (refer to test data), presence of 15 or more nucleotides in predicted promoter sequences (75 bp window) were considered successful predictions. Any predictions with less than 15 nucleotides of the original 101 bp fragments were considered incorrect predictions.

6.2.2. Test Data

6.2.2.1. First Test Data (fragment sizes of 481 bp).

Eighty (80) of the original 83 test promoter sequences from both *E.coli* and *B.subtilis* test promoters were located in their respective annotated genomes. Each promoter sequence (101 bp) was extracted together with additional 190 bp on either side of the promoter in the genome. Eighty, instead of the original eighty-three (83) *E.coli* promoter test data were used because three promoters in the original *E.coli* promoter dataset from Lisser and Margalit (1993) could not be located in the annotated genome data (ecoli.fna). The three promoters that were not found in the annotated *E.coli* genome have been documented in Appendix_eleven. Because the three promoters from *E.coli* were not available, eighty (80) *B.subtilis* promoters were used instead of eighty-three (to maintain uniformity on test data with that used for *E.coli*). The total fragment size of each test sequence came up to 481 bp. Test datasets used for *E.coli* and *B.subtilis* can be found in Appendix_twelve and Appendix_thirteen respectively.

UNIVERSITY of the
WESTERN CAPE

6.2.2.2. Second Test Data (sections covering ~5-10 kb of genome)

Individual genomes of both *E.coli* and *B.subtilis* were scanned for regions that had at least three of the test promoter sequences of respective organisms within a section of ~10 kb nucleotides. Two such regions covering approximately 6 kb for *B.subtilis* (Appendix_fifteen) and 11 kb for *E.coli* (Appendix_fourteen) were selected. The selected *B.subtilis* region contained known promoters *veg* (vegetative), *sspF* (small acid-soluble spore proteins) and *spoVG* (sporulation) whilst *E.coli* region harbored promoters *aroP* (a member of the *tyrR* regulon), *aceE* (pyruvate dehydrogenase complex) and *lpd* (pyruvate dehydrogenase complex).

6.2.2.3. Third Test Data (Regions upstream of annotated genes)

Annotated genome files of *E.coli* (ecoli.ffn and ecoli.fna), *B.subtilis* (bsub.ffn and bsub.fna) and *M.tuberculosis* (mtub.ffn and mtub.fna) were obtained from Genbank (version 111). Using the respective annotated genome files (extension ffn), regions between the coding sequences were processed as intergenic regions. Annotated genes were classified into four groups using next consecutive genes.

Category A: gene in direct frame followed by another in direct frame (direct and parallel).

Category B: gene in direct frame followed by complement gene (convergent).

Category C: complement gene followed by another complement gene (complementary and parallel).

Category D: complement gene followed by a gene in direct frame (divergent).

Nucleotide sequences between convergent genes were ignored since these region promoters are not supposed to harbor promoters. Category D genes (divergent) were extracted and processed for two promoters in respective orientations. All sequences between category A and category C genes were extracted. On relatively few instances where category D genes overlapped, 250 bp nucleotides upstream of the genes were extracted and used as inter-orfs.

WESTERN CAPE

6.2.3. Types of Test.

Types of tests carried on the test datasets listed above are as follows:

A) Given a fragment of sequence (481 bp for first test data and inter-orf region for third test data), the sequence fragment (75 bp window) with the best score for each prediction method was selected as the predicted promoter in the entire section. Thus, three sets of prediction results were generated (from the three different methods) for each of the eighty test datasets.

B) Combined predictions from the all the three methods on first and third test data sets. Initial prediction values (all three systems) for the first 75 bp window were stored together with the sequence in the window (75 bp). Subsequent predictions as the test window was moved in 1 bp increments were compared to the previous ones. Predicted sequence was replaced only when all three scores were better than previous corresponding scores. Thus only one prediction score for each test case was generated.

(C) About 5 to 10 kb of nucleotides (with three known promoters) covering sections of respective genomes of both *E.coli* and *B.subtilis* (second test data) were analyzed using a 75 bp window. Prediction results (all three methods) of each 75 bp window was recorded and plotted on graph, as the window was shifted in 1 bp increments to the end. The plot is meant to portray the 'signals' as one 'walks' along a section of the genomes of the respective organisms.

UNIVERSITY of the
WESTERN CAPE

6.2.4. Choice of Models for Integrated Prediction

The models/profiles that produced the best results in the 75 bp test categories (test type B) for all the three prediction systems namely, Hidden Markov Model, Neural Network and TFDA in the previous three chapters are shown below.

	E.coli	*B.subtilis*	Mycobacterium
HMM	50_45	50_75	30_75
NN	30_60	30_55	50_60
TFDA	30_60	50_50	40_70

Figure 6.1. The various models reflecting sequence subsets that produced best results in the 75 bp test category (type B) for the three prediction systems in the three organisms. As denoted earlier, 50_45 represents a sequence subset comprising 50 sequences of 45 bp fragment sizes each. The models developed on these subsequences were used in all the predictions in this chapter.

6.2.5. Prediction Quality Ranking Methods

A simple system of ranking based on results of predictions from all three prediction methods on the eighty test sequences (section 6.2.3) was used in the combined prediction tool. Comparison of predicted scores to previous predictions were in the order TFDA, HMM and NN for both *E.coli* and *B.subtilis*. This is based on the results represented in figure 6.2. For example, TFDA results were compared first because TFDA had the highest score of predicted results followed by HMM in *B.subtilis* though both ANN and HMM scored same number of correct predictions in *E.coli*.

UNIVERSITY of the
WESTERN CAPE

6.3. Results and Discussion

Prediction results of the individual methods on first data set (a promoter lying between 190 bp nucleotides on either side) for both *E.coli* and *B.subtilis* are shown in figure 6.2A and 6.2B respectively. Models trained on subsets of 50_45, 30_60 and 30_60 representing prediction methods HMM, NN and TFDA respectively were used entirely in this chapter. Promoter predictions were on a 75 bp window. A prediction was categorized as positive if more than a 15 bp section of the original promoter was found in the predicted 75 bp window. Of the 80 promoter test sequences, 30 *E.coli* promoters were correctly predicted by the three prediction systems in *E.coli* whilst forty-four (44) were correctly predicted on *B.subtilis* testdata. In almost all the thirty predictions on *E.coli*, predicted sequences covered more than 55 bp of original promoters (Appendix_sixteen). Analysis of the predicted results on *E.coli* test data uncovered TFDA as the best prediction method of the three (HMM, ANN and TFDA). Prediction from TFDA (47/80) produced five (5) more positives than results from both HMM and NN (42/80 each). In all, 72.5% (58/80) of the *E.coli* test data (first test data) were correctly predicted by at least one of the prediction models and therefore there existed a 27.5% false positive rate.

A**.** *E.coli*

Total Number of test sequences = 80

Size of each test fragment = 481

Test sequences **not predicted** by any of the **three** =22

Test sequences predicted correctly by all three (3) = 30

Test sequences predicted correctly by HMM = 42

Test sequences predicted correctly by ANN = 42

Test sequences predicted correctly by TFDA = 47

UNIVERSITY of the
WESTERN CAPE

B. *B.subtilis*

Total Number of test sequences = 80

Size of each test fragment = 481

Test sequences **not predicted** by any of the **three** =9

Test sequences predicted correctly by all three (3) = 44

Test sequences predicted correctly by HMM = 61

Test sequences predicted correctly by ANN = 55

Test sequences predicted correctly by TFDA = 62

Figure 6.2. Prediction results on *E.coli* (A) and *B.subtilis* (B) using the subset models of the three prediction methods (figure 6.1). Test data consisted of 80

genome sequences each of 481 bp fragment sizes (first test data). Results are the best predictions from the individual models (Appendix_sixteen).

UNIVERSITY of the
WESTERN CAPE

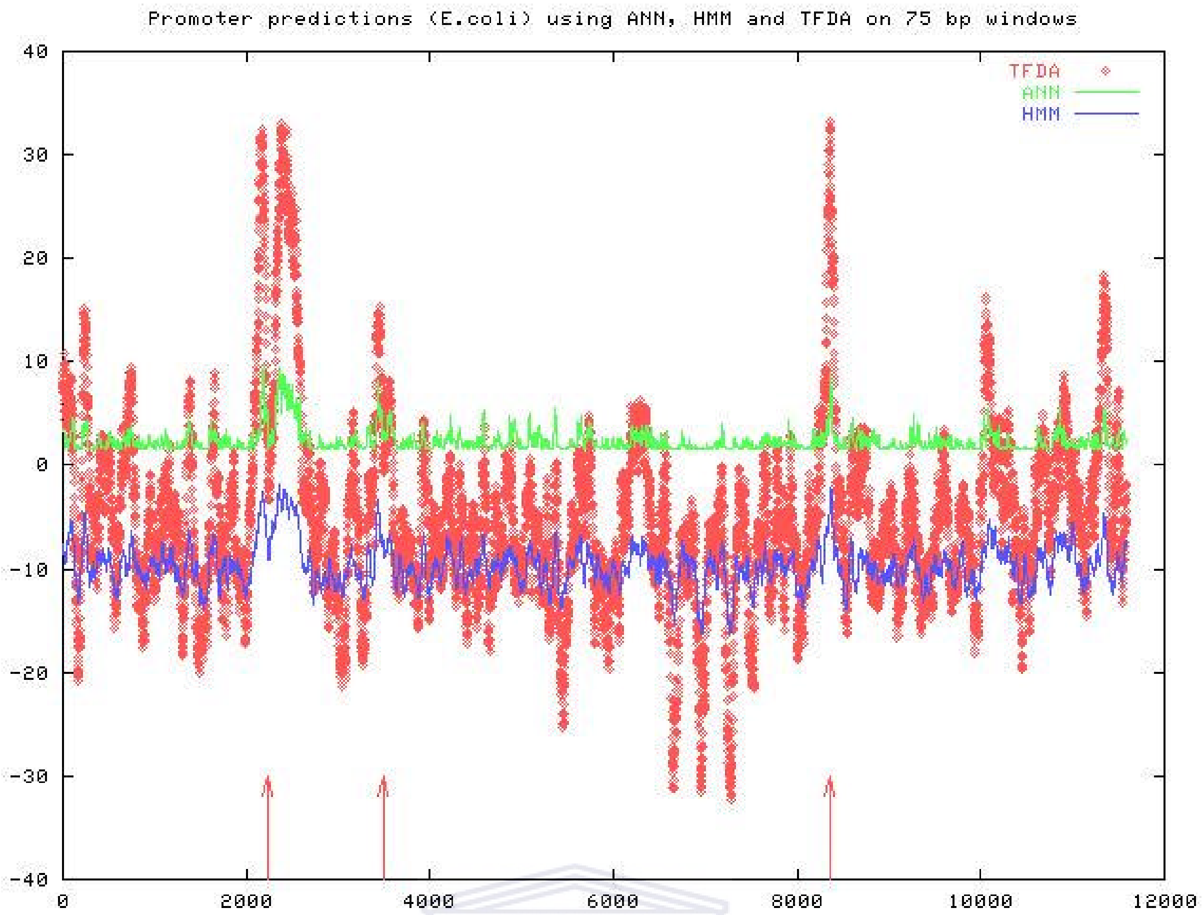


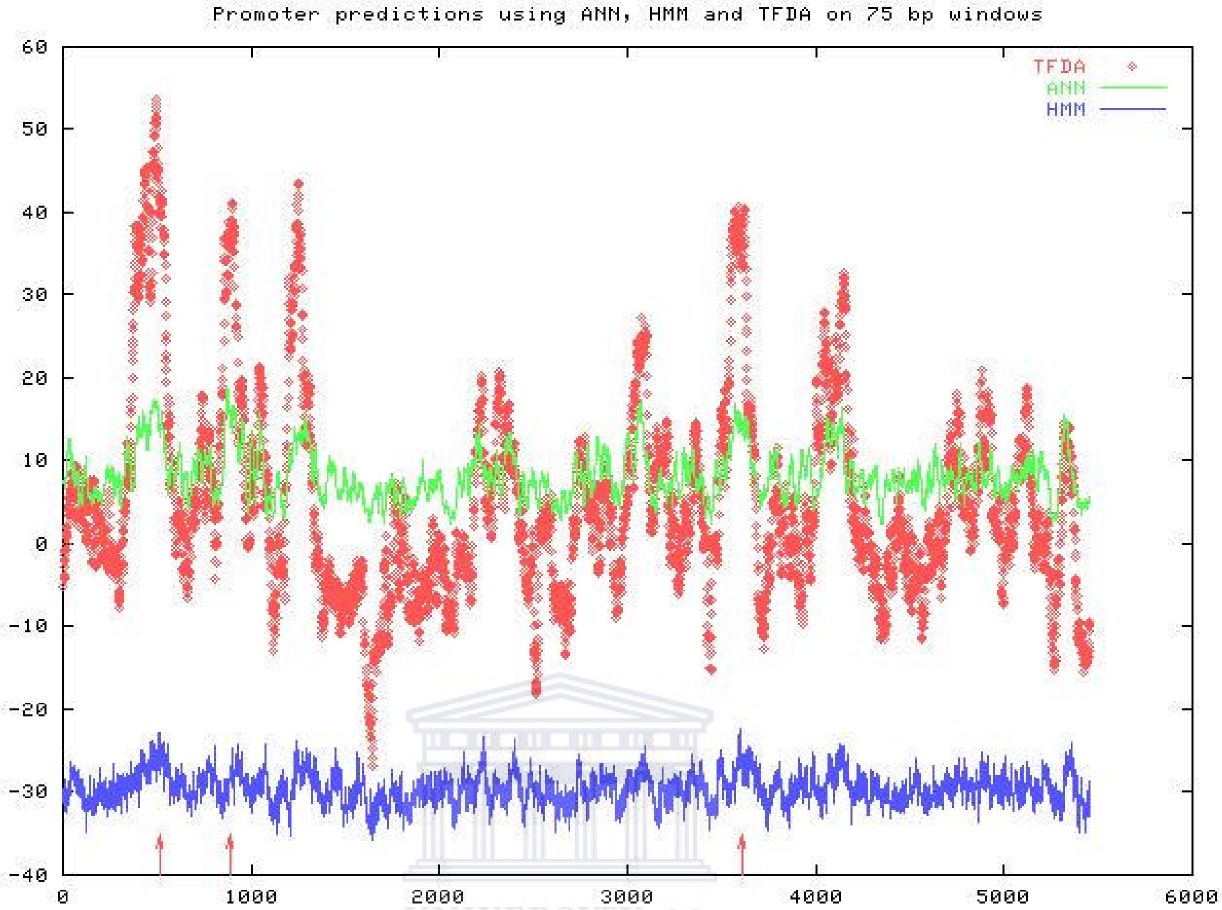
Figure 6.3. Prediction results on a section of E.coli genome harboring promoters *aroP*, *aceE* and *lpd*. A 75-bp window was used for predictions. Scores on HMM, ANN and TFDA were adjusted to accommodate all three on the same plot. Results from prediction were obtained by continuously moving the window one bp till the end of the sequence. Positions of the three promoters namely *aroP*, *aceE* and *lpd* in the dataset are represented by the arrows at positions 2226, 3493 and 8362 respectively. Individual predictions from the three separate methods ANN, HMM and TFDA on the same test data can be found at Appendix_twenty, Appendix_twentyone and Appendix_twentytwo respectively.

UNIVERSITY of the
WESTERN CAPE

Prediction results from *B.subtilis* datasets on the similar test data (first test data) were much better, with 42 predictions from all the three prediction systems and only nine (9) ‘incorrect’ predictions (Appendix_seventeen). Seventy-one (71) of the eighty (80) test data were correctly predicted by at least one of the methods. Once again, TFDA excelled as the most efficient predicting system for *B.subtilis* with 62 correct predictions of the 80 promoters correctly predicted. Correct predictions for HMM, NN and TFDA were 61, 55 and 62 respectively (Appendix_seventeen). The second sets of predictions were performed on the same testdata (first test data), this time combining the strengths of the three prediction methods. Prediction results had to be the best from the combination of all three predictions as explained in the section 6.2.3B. Results from the combined predictions on both *E.coli* and *B.subtilis* can be found in Appendix_eighteen and Appendix_nineteen respectively. Thirty-three (33) of *E.coli* promoters were not predicted ‘correctly’ as compared to fifteen (15) *B.subtilis* promoters.

The test results on sequences covering a region around promoters *aroP*, *aceE* and *lpdA* for *E.coli* and *veg*, *sspF* and *spoVG* for *B.subtilis* are shown in figures 6.3 and 6.4 respectively. Peaks are evident at sections where the known promoter sequences were located as indicated by arrows in the respective diagrams. Comment is reserved on the other peaks as the study was focused on the known promoters. Individual plots of the prediction system can be found in Appendix_twenty to Appendix_twentytwo for *E.coli* and Appendix_twentythree to Appendix_twentyfive for *B.subtilis*. Of particular interest is the second peak around 2500 after the *aroP* promoter’s peak (2226) in the *E.coli* data. The peak (~2500) in question portrays the sequence window (~75 bp) covering the region to promoter features. The peak is reflected in the plots of all the three prediction systems (figure 6.3). Sequence windows portraying such peaks need to be experimentally analyzed for promoter function(s). Such an analysis would give an indication of what to expect from similar peaks throughout the genome.

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 6.4. Prediction scores of NN (green), HMM (blue) and TFDA (red) on 75 bp window sized sequences covering ~5500 bp region of *B.subtilis* genome harboring promoters *veg*, *sspF* and *spoVG*. Test sequences and prediction scores were obtained by shifting each previous window by 1 bp. Results from HMM were multiplied by (0.35) to enable the values to fit onto the graphs. Promoters *veg*, *ssPf* and *spoVG* are found in positions 520, 890 and 3606 respectively as indicated by the arrows. The individual plots for predictions of ANN, HMM and TFDA can be found in Appendix_twentythree, Appendix_twentyfour and Appendix_twentyfive respectively.

Promoter test sequence predictions that were 'incorrect' were investigated against the background of the predicted promoter sequences. The predicted sequences obviously have more promoter features/characteristics than the original promoter sequences in the test data. In most of the 'incorrect' predictions for *E.coli*, a least two of the three predictions were on specific 75-bp window. Most of the 'incorrect' predictions had all three predictions (*uvrA*, *purD, malP, pnp, sulA, and pncB* among others) within the same 75 bp window (Appendix_sixteen). Four possible explanations come to mind with such predictions. (1) The predicted promoters may be second/alternative promoters that have not been established. (2) The predicted promoters may have their functions suppressed by other cis-acting sequences, for example, oppressors. (3) The actual promoters not predicted may be used by different or alternative sigma factors. (4). Finally, the features analyzed in the study are not sufficiently characteristic of functional promoters. These are of course hypotheses as they have not been proven experimentally. Whatever the case is, the 'incorrect' predictions noted as false positives cannot just be ignored.

Dual or multiple promoters have been found in *E.coli* (*uvrB*, van den Berg *et al*., 1983; *pheV,* Caillet *et al*., 1985*; metY-nusA-infB*, Granston *et al*., 1990;), *B.subtilis* (*spoVG*, Johnson *et al*., 1983; *opuE*, Spiegelhalter and Bremer, 1996) and *M.tuberculosis* (*recA*, Mohahedzadeh *et al*., 1997; *katG*, Andesen *et al*., 1988). It is interesting to note that, almost all the twenty-two 'incorrectly' predicted *E.coli* promoters, figure 6.2A are promoters of house keeping genes. The 'incorrectly' predicted *E.coli* promoters are indexed in Appendix_twentysix. Perhaps, the predicted promoters are alternative promoters with an unknown role(s). Certainly such an occurrence would be fatal to the organism. Still, why the actual promoters are preferred by RNA polymerase and the associating sigma factors to the predicted promoters is an interesting phenomenon that needs to be investigated further. The nine *B.subtilis* promoter sequences not predicted by any of the three prediction systems can be found in Appendix_twentyseven.

A slightly different picture is observed with *B.subtilis* where ~89% (71/80) percent of the promoters were correctly predicted by at least one of the prediction systems. Of the nine 'incorrectly' predicted promoters namely *spoVE, rpoD, degQ, cotF,* *cspB, cotH, spoIIIG, cotB* and *abr* (Appendix_twentyseven), seven of the predictions were centered around the same 75 bp window in all the three predictions for each predicted promoter. Only predictions for promoters of *abr* and *cotH* did not have the three predictions not covering a region around the same 75 bp window. Even then, for *cotH*, both HMM and ANN predictions were similar (covered about the same 75 bp window). Similarly, HMM and TFDA predicted the almost the same window for *abr* promoter. The prediction results for *spoVE* for example centered on the first promoter P1 of its tandem promoters P1 and P2 (Miyao *et al*., 1993). Thus, it is very important to incorporate graphical information in analyzing these predictions so that the predictions can be viewed in context to the neighboring nucleotide sequences. Another important reason for incorporating graphical information is the fact that, predicted results were relative to the test sequences. Whereas for example, a TFDA score of 6.3000 for a window may be the highest in a particular test sequence, it may not probably appear in the top ten of another test sequence. In the application of the prediction system to the intergenic regions of entire genomes, sequences less than 75 bp were left untouched. Sequences from 75 to about 200 represented a relatively easy task with respect to correct predictions being made. The problematic area could be with intergenic sequences over 400 bp. In *E.coli* for example, over 330 intergenic regions were found to be nucleotides greater than 400 bp. The results on entire intergenic regions of *E.coli*, *B.subtilis* and *M.tuberculosis* can be found in Appendix_twentyeight, Appendix_twentynine and Appendix_thirty respectively. Since only the regions between genes are analyzed, promoter sequences located in coding sequences would be missed. Sadly, little can be done at this stage until more information is available on prokaryotic transcription machinery. The current approach is; promoter sequences must be referenced to the genes or operons and therefore focus has been on the inter-orfs. It is therefore envisaged that, users of the predicted information will study the graphical predictions around the immediate region surrounding each predicted promoter and not take the predicted promoter out of context with the surrounding sequences. Though true positive scores are relatively high even among 481 bp sequences, some promoters used by specific RNA

UNIVERSITY of the
WESTERN CAPE

polymerases may be missed in the prediction. This may not necessarily be false or non-promoters. Possible explanations include probable alternate promoters or promoters being used by or co-transcribed alternative sigma factors.

Chapter seven

Conclusion

Promoter detection, especially in prokaryotes, has always been an uphill task and may remain so, because of the many varieties of sigma factors employed by various organisms in transcription. The situation is made more complex by the fact, that any seemingly unimportant sequence segment may be turned into a promoter sequence by an activator or repressor (if the actual promoter sequence is made unavailable). Nevertheless, a computational approach to promoter detection has to be performed due to number of reasons. The obvious that comes to mind is the long and tedious process involved in elucidating promoters in the ‘wet’ laboratories not to mention the financial aspect of such endeavors. Promoter detection/prediction of an organism with few characterized promoters (*M.tuberculosis*) as envisaged at the beginning of this work was never going to be easy. Even for the few known Mycobacterial promoters, most of the respective sigma factors associated with their transcription were not known. If the information (promoter-sigma) were available, the research would have been focused on categorizing the promoters according to sigma factors and training the methods on the respective categories. That is assuming that, there would be enough training data for the respective categories. Most promoter detection/prediction studies have been carried out on *E.coli* because of the availability of a number of experimentally characterized promoters (+- 310). Even then, no researcher to date has extended the research to the entire *E.coli* genome.

UNIVERSITY of the WESTERN CAPE

Since prokaryotic promoter detection in various forms have been tackled by other researchers using various methods, the idea of integrating some prediction methods using small training data have been very appealing. Plus, with the recent advancement in genome sequencing techniques, there will always be a lot of annotated (genes) available. Annotation is really incomplete without other chromosomal features such as promoters, oppressors and repressors, thus reinforcing the need to tackle the promoter detection problem.

We have used promoter sequences available for *E.coli*, *B.subtilis* and *M.tuberculosis* and trained models of ANN, HMM and TFDA (creation of author) on these promoters to study promoter detection and prediction. The study has resulted in creation of database of predicted promoters of *E.coli* and *B.subtilis* at the website of South African National Bioinformatics (SANBI) website. Experience gained on the study on *E.coli* and *B.subtilis* have been applied to establish a similar database for *M.tuberculosis* at the same website. Three types of information on the promoters are available at the website for researchers:

1. The best predicted promoter sequence for particular genes or operons (in a 75 bp size window using the combined strength of all the three prediction systems.
2. The best predictions from the individual prediction systems, that is best of ANN, HMM and TFDA for any particular gene or operon.
3. A chance to have a graphical view of the prediction scores from all three prediction systems on the sequence region of interest.

With such information available, and with information like –10 and –35 hexamers, ribosomal binding sites which were not directly incorporated in the study, we are certain that researchers will be able to eyeball the promoters of their respective genes under investigations if they are not known. A useful exercise will be to pick about ten of the *M.tuberculosis* predicted promoters randomly from the database and test their ability to enable the transcription of the respective genes.

UNIVERSITY of the
WESTERN CAPE

APPENDICES

Appendices have been placed at the following ftp site.

ftp.sanbi.ac.za/pub/ekow/APPENDIX.

Username: anonymous

Password: email address

Or

ftp://ftp.sanbi.ac.za/pub/ekow/APPENDIX

UNIVERSITY of the
WESTERN CAPE

Postscript files may have to be downloaded and viewed with appropriate postscript viewer such as gsview, psview or ghostview.

References

Andersen, A.B., Worsaae, A. and Chaparas, S.D. (1988). Isolation and characterization of recombinant lambda gt11 bacteriophages expressing eight different mycobacterial antigens of potential immunological relevance. Infect. Immun. May; 56(5): 1344-51.

Asai, K., Hayamizu, S. and Handa, K. (1993). Prediction of protein secondary structure by the hidden Markov model. Comput. Appl. Biosci. Apr;9(2):141-6.

Asai, K. (1996). Extraction of hidden Markov model representations of signal patterns in DNA sequences. Pac. Symp. Biocomput. 686-96.

Baker, R.E., Camier, S., Sentenac, A. and Hall, B.D. (1987). Gene size differentially affects the binding of yeast transcription factor tau to two intragenic regions. Proc. Natl. Acad. Sci. (U SA) Dec;84(24):8768-72.

Bairoch, A. and Bucher, P. (1994). PROSITE: recent developments. Nucleic Acids Res. Sep;22(17):3583-9.

Baldi, P. and Chauvin, Y. (1995). Protein modeling with hybrid Hidden Markov Model/Neural network architectures. Ismb. 3:39-47.

Baldi, P., Brunak, S., Chauvin, Y., Engelbrecht, J. and Krogh, A. (1995). Periodic sequence patterns in human exons. Ismb. 3:30-8.

Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M.A. (1994). Hidden Markov models of biological primary sequence information. Proc. Natl. Acad. Sci. U.S.A. Feb. 1;91(3):1059-63.

WESTERN CAPE

Barrett, C., Hughey, R. and Karplus, K. (1997). Scoring hidden Markov models. Comput. Appl. Biosci. Apr;13(2):191-9.

Barton, G.J. (1990). Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol. 183:403-28.

Bashyam, M.D., Kaushal, D., Dasgupta, S.K., Tyagi, A.K. (1996). A study of mycobacterial transcriptional apparatus: identification of novel features in promoter elements. J. Bacteriol. Aug;178(16):4847-53.

Bark, C., Weller, P., Zabielski, J., Janson, L. and Pettersson, U. (1987). A distant enhancer I element is required for polymerase III transcription of a U6 RNA gene. Nature. Jul 23-29;328(6128):356-9.

Bell, S.P., Pikaard, C.S., Reeder, R.H. and Tjian, R. (1989). Molecular mechanisms governing species-specific transcription of ribosomal RNA. Cell Nov 3;59(3):489-97.

Bell, S.P., Jantzen, H.M. and Tjian, R. (1990). Assembly of alternative multiprotein complexes directs rRNA promoter selectivity. Genes Dev. Jun;4(6):943-54.

WESTERN CAPE

Belyaeva, T., Griffiths, L., Minchin, S., Cole, J. and Busby, S. (1993). The *Escherichia coli cysG* promoter, the 'extended -10' class of bacterial promoters. Biochem. J. Dec. 15;296 (Pt 3):851-7.

Besemer, J. and Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. Nucleic Acids Res. Oct 1;27(19):3911-20.

Bourn, W.R. and Babb, B. (1995). Computer assisted identification and classification of *streptomycete* promoters. Nucleic Acids Res. 23(18):3696-703.

Borodovsky, M. McIninch, J.D., Koonin, E.V., Rudd, K.E. Medigue, C. and Danchin, A.(1995). Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res. Sep11;23(17):3554-62.

Borodovsky, M. and McIninch, J. (1993). GeneMark: Parallel Gene Recognition for both DNA Strands. Computers & Chemistry, 17, 123-133.

Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. J. Mol. Biol. 220(1):49-65.

Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. Comput. Chem. Mar;20(1):3-23.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. Apr 25;268(1):78-94.

Burns, H. and Minchin, S. (1994). Thermal energy requirement for strand separation during transcription initiation: the effect of supercoiling and extended protein DNA contacts. Nucleic Acids Res. Sep. 25;22(19):3840-5.

Busby, S. and Ebright, R.H. (1994). Promoter structure, promoter recognition and transcription activation in prokaryotes. Cell Dec 2;79(5):743-6.

Caillet, J., Plumbridge, J.A. and Springer, M. (1985). Evidence that *pheV*, a gene for *tRNAPhe* of *E.coli* is transcribed from tandem promoters. Nucleic Acids Res. May 24;13(10):3699-71.

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(1):159-70.

UNIVERSITY of the
WESTERN CAPE

Chamberlin, M.J. (1974). The selectivity of transcription. Annu. Rev. Biochem. 43(0):721-75.

Chan, B. and Busby, S. (1989). Recognition of nucleotide sequences at the *Escherichia coli* galactose operon P1 promoter by RNA polymerase. Gene, Dec.14;84(2):227-36.

Choo, Y., Castellanos, A., Garcia-Hernandez, B., Sanchez-Garcia, I. and Klug, A. (1997). Promoter-specific activation directed by bacteriophage-selected zinc fingers. J. Mol. Biol. Oct 31;273(3):525-32.

Das Gupta S.K, Bashyam M.D, Tyagi A.K (1993). Cloning and assessment of mycobacterial promoters by using a plasmid shuttle vector. J. Bacteriol. Aug;175(16):5186-92.

Dellagostin, O.A., Esposito, G., Eales, L.J., Dale, J.W. and McFadden, J. (1995). Activity of mycobacterial promoters during intracellular and extracellular growth. Microbiology Aug;141 (Pt 8):1785-92.

UNIVERSITY of the
WESTERN CAPE

Dhandayuthapani, S., Zhang, Y., Mudd, M.H. and Deretic, V. (1996). Oxidative stress response and its role in sensitivity to isoniazid in mycobacteria: characterization and inducibility of *ahpC* by peroxides in *Mycobacterium smegmatis* and lack of expression in *M.aurum* and *M. tuberculosis*. J. Bacteriol. Jun;178(12):3641-9.

Demeler, B. and Zhou, G.W. (1991). Neural network optimization for *E.coli* promoter prediction. Nucleic Acids Res. 19(7):1593-9.

Deuschle, U., Kammerer, W., Gentz, R. and Bujard, H. (1986). Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. EMBO J Nov;5(11):2987-94.

Eddy, S.R. (1995). Multiple alignment using hidden Markov models. Ismb. 3:114-20.

Eddy, S.R. (1996). Hidden Markov models. Curr. Opin. Struct. Biol. Jun;6(3):361-5.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics 14(9):755-63.

Eddy, S.R. Mitchison, G. and Durbin, R. (1995). Maximum discrimination hidden Markov models. J. Comput. Biol. Spring;2(1):9-23.

Fabrizio, P., Coppo, A., Fruscoloni, P., Benedetti, P. and Di Segni, G. and Tocchini-Valentini, G.P. (1987). Comparative mutational analysis of wild-type and stretched tRNA3(Leu) gene promoters. Proc. Natl. Acad. Sci. (U S A) Dec;84(24):8763-7.

Felsenstein, J. and Churchill, G.A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. Jan;13(1):93-104.

Fickett, J.W. (1998). Predictive methods using nucleotide sequences. Methods Biochem. Anal;39:231-45.

Firek, S., Read, C., Smith, D.R. and Moss, T. (1990). Point mutation analysis of the *Xenopus laevis* RNA polymerase I core promoter. Nucleic Acids Res. Jan 11;18(1):105-9.

Frishman D., Mironov, A., Mewes, H.W. and Gelfand, M.(1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res., 26, 2941-2947.

Galas, D.J., Waterman, M.S. and Arratia. R. (1984). Pattern recognition in several sequences: consensus and alignment. Bull Math Biol. 46(4):515-27.

UNIVERSITY of the
WESTERN CAPE

Gatlin, L.L. (1966). The information content of DNA. J. Theor. Biol. Feb;10(2):281-300.

Goldman, N., Thorne, J.L. and Jones, D.T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. Oct. 25;263(2):196-208.

Goodrich, J.A. and McClure, W.R. (1992). Regulation of open complex formation at the *Escherichia coli* galactose operon promoters. Simultaneous interaction of RNA polymerase *gal* repressor and CAP/cAMP. J. Mol. Biol. Mar. 5;224(1):15-29.

Goodrich, J.A., Cutler, G. and Tjian R (1996). Contacts in context: promoter specificity and macromolecular interactions in transcription. Cell, Mar 22;84(6):825-30.

Granston, A.E., Thompson, D.L. and Friedman, D.I. (1990). Identification of a second promoter for the metY-nusA-infB operon of *Escherichia coli*. J. Bacteriol. May;172(5):2336-42.

Graves, M.C. and Rabinowitz, J.C. (1986). In vivo and in vitro transcription of the *Clostridium pasteurianum ferredoxin* gene. Evidence for "extended" promoter elements in gram-positive organisms. J. Biol. Chem. Aug 25;261(24):11409-15.

Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. Jul;84(13):4355-8.

Grummt, I., Roth, E. and Paule, M.R. (1982). Ribosomal RNA transcription in vitro is species specific. Nature Mar. 11;296(5853):173-4.

Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992). Prediction of gene structure. J. Mol. Biol; Jul 5;226(1):141-57.

UNIVERSITY of the
WESTERN CAPE

Harley, C.B. and Reynolds, R.P. (1987). Analysis of *E.coli* promoter sequences. Nucleic Acids Res. Mar 11;15(5):2343-61.

Haussler, D., Brown, M., Hughey, R., Krogh, A., Mian, I.S. and Sjolander, K. (1993). Using Dirichlet mixture priors to derive hidden Markov models for protein families. Ismb.1:47-55.

Hawley, D. K., McClure, W.R. (1982). Mechanism of activation of transcription initiation from the lambda PRM promoter. J. Mol. Biol., 157(3):493-525.

Hawley, D.K. and McClure, W.R. (1983). The effect of a lambda repressor mutation on the activation of transcription initiation from the lambda PRM promoter. Cell Feb;32(2):327-33.

Hawley, D.K and McClure, W.R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res. Apr 25;11(8):2237-55.

Helmann, J.D. (1995) Compilation and analysis of *Bacillus subtilis* sigma A-dependent Promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter and DNA. Nucleic Acids Res. Jul. 11;23(13):2351-60.

Henderson, S.L. and Sollner-Webb, B. (1990). The mouse ribosomal DNA promoter has more stringent requirements in vivo than in vitro. Mol Cell Biol Sep;10(9):4970-3.

Henderson, J., Salzberg, S. and Fasman, K.H. (1997). Finding genes in DNA with a Hidden Markov Model. J. Comput. Biol. Summer;4(2):127-41.

Henikoff, S. (1996). Scores for sequence searches and alignments. Curr. Opin. Struct. Biol. Jun;6(3):353-60.

UNIVERSITY of the WESTERN CAPE

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K. and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. Science. Oct 24;278(5338):609-14.

Henkin, T.M. and Sonenshein, A.L. (1987). Mutations of the *Escherichia coli lacUV* promoter resulting in increased expression in *Bacillus subtilis*. Mol Gen. Genet. Oct;209(3):467-74.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. Apr;79(8):2554-8.

Horton, P.B. and Kanehisa, M. (1992). An assessment of neural network and statistical approaches for prediction of *E.coli* promoter sites. Nucleic Acids Res. 20(16):4331-8.

Hughey, R. and Krogh, A. (1996). Hidden Markov models for sequence analysis: Comput.Appl. Biosci. Apr;12(2):95-107.

UNIVERSITY of the
WESTERN CAPE

Hutchinson, G. B. (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. Comput. Appl. Biosci. Oct;12(5):391-8.

Karchin, R and Hughey, R. (1998). Weighting hidden Markov models for maximum discrimination. Bioinformatics,14(9):772-82.

Kaufmann, J. and Smale, S.T. (1994). Direct recognition of initiator elements by a component of the transcription factor IID complex.

Klug and Cummings (1997). Concept of Genetics (Fifth Edition) Prentice Hall International. Ed.

Knudsen, S. (1999). Promoter2.0: for the recognition of PolIII promoter sequences. Bioinformatics May;15(5):356-61.

Janssen, G.R. and Bibb, M.J. (1990). Tandem promoters, tsrp1 and tsrp2, direct transcription of the thiostrepton resistance gene (tsr) of *Streptomyces azureus*: transcriptional initiation from tsrp2 occurs after deletion of the -35 region. Mol. Gen. Genet. May;221(3):339-46.

Johnson, W.C., Moran, C.P. Jr. and Losick, R. (1983). Two RNA polymerase sigma factors from *Bacillus subtilis* discriminate between *Bacillus subtilis* discriminate between regulated gene. Nature Apr 28;302(5911):800-4.

Kahl, B.F., Li, H. and Paule, M.R. (2000). DNA melting and promoter clearance by eukaryotic RNA polymerase I. J. Mol. Biol. May 26;299(1):75-89.

Keilty, R. and Rosenberg, M. (1987). Constitutive function of a positively regulated promoter
reveals new sequences essential for activity. J. Biol. Chem. May 5;262(13):6389-95.

Kenney, T.J. and Churchward, G. (1996). Genetic analysis of the *Mycobacterium smegmatis rpsL* promoter. J. Bacteriol. Jun;178(12):3564-71.
Kobayashi, M. Nagata, K. and Ishihama, A. (1990). Promoter selectivity of *Escherichia coli* RNA polymerase: effect of base substitutions in the promoter -35 region on promoter strength. Nucleic Acids Res. Dec 25;18(24):7367-72.

Kownin, P., Iida, C.T., Brown-Shimer, S. and Paule, M.R. (1985). The ribosomal RNA promoter of *Acanthamoeba castellanii* determined by transcription in a cell-free system. Nucleic Acids Res. Sep 11;13(17)6237-6248.

WESTERN CAPE

Kremer, L., Baulard, A., Estaquier, J., Content, J., Capron, A. and Locht, C. (1995). Analysis of the *Mycobacterium tuberculosis* 85A antigen promoter region. J. Bacteriol Feb;177(3):642-53.

Krogh, A. (2000). Using database matches with for HMMGene for automated gene detection in Drosophila. Genome Res. Apr;10(4):523-8.

Krogh, A., Mian, I.S. and Haussler, D. (1994). A hidden Markov model that finds genes in *E. coli* DNA. Nucleic Acids Res. Nov 11;22(22):4768-78.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. 1996 Jun;58(6):1347-63.

Kulp, D., Reese, M.G. and Eeckman, F.H. (1996). A Generalized Hidden Markov Model for the Recognition of Human genes in DNA. Ismb 96: 134-142.

Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1997). Integrating database homology in a probabilistic gene structure model. Pac. Symp. Biocomput:232-44.

Laalami, S., Timofeev, A.V., Putzer, H., Leautey, J. and Grunberg-Manago, M. (1994). In vivo study of engineered G-domain mutants of *Escherichia coli* translation initiation factor IF2. Mol Microbiol Jan;11(2):293-302.

Lazareva-Ulitsky, B. and Haussler, D. (1999). A probabilistic approach to consensus multiple alignment. Pac. Symp. Biocomput. 150-61.

Lee, Y., Erkine, A.M., Van Ryk, D.I. and Nazar, R.N. (1995). In vivo analyses of the internal control region in the 5S rRNA gene from *Saccharomyces cerevisiae*. Nucleic Acids Res. Feb 25;23(4):634-40.

UNIVERSITY of the
WESTERN CAPE

Lewin, B. (1997) Genes VI. Oxford University Press. 287-332.

Levin, M.E. and Hatfull, G.F. (1993). Mycobacterium smegmatis RNA polymerase: DNA supercoiling, action of rifampicin and mechanism of rifampicin resistance. Mol. Microbiol. Apr;8(2):277-85.

Lisser, S. and Margalit, H. (1993) Compilation of *E.coli* mRNA promoter sequences. Nucleic Acids Res. 21(7):1507-16.

Lobo, S.M. and Hernandez, N. (1989). A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter. Cell Jul 14;58(1):55-67.

Lukashin, A.V., Anshelevich, V.V., Amirikyan, B.R., Gragerov, A.I. and Frank-Kamenetskii M.D. (1989). Neural network models for promoter recognition. J. Biomol. Struct. Dyn. 6:1123-33.

Lukashin, A.V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for genefinding. Nucleic Acids Res. Feb 15;26(4):1107-15.

Mahadevan, I. and Ghosh, I. (1994) Analysis of *E.coli* promoter structures using neural networks. Nucleic Acids Res. 22(11):2158-65.

Mamitsuka, H (1996). A learning method of hidden Markov models for sequence discrimination. J. Comput. Biol. Fall;3(3):361-73.

Mangel, W.F. and Chamberlin, M.J. (1974) Studies of ribonucleic acid chain initiation by *Escherichia coli* ribonucleic acid polymerase bound to T7 deoxyribonucleic acid. 3. The effect of temperature on ribonucleic acid chain initiation and on the conformation of binary complexes. J. Biol. Chem. 249(10):3007-13.

Marilley, M. and Pasero, P. (1996). Common DNA structural features exhibited by eukaryotic ribosomal gene promoters. Nucleic Acids Res. Jun 15;24(12):2204-11.

McClure, M.A., Smith, C. and Elton, P. (1996). Parameterization studies for the SAM and HMMER methods of hidden Markov model generation. Ismb 4:155-64.

Miyao, A., Theeragool, G., Takeuchi, M. and Kobayashi, Y. (1993). *Bacillus subtilis spoVE* gene is transcribed by sigma E-associated RNA polymerase. J. Bacteriol.Jul;175(13):4081-6.

Moran, C.P. Jr., Johnson, W.C. and Losick, R. (1982). Close contacts between sigma 37-RNA polymerase and a *Bacillus subtilis* chromosomal promoter. J. Mol. Biol. Dec 15;162(3):709-13.

Moss, B and Jones E.V., (1984). Transcriptional mapping of the vaccinia virus DNA polymerase gene. J.Virol. Jan;53(1):312-5.

Moss, T. and Stefanovsky, V.Y. (1995). Promotion and regulation of ribosomal transcription in eukaryotes by RNA polymerase. Prog. Nucleic Acid Res. Mol. Biol.50:25-66.

Movahedzadeh, F., Colston, M.J. and Davis, E.O. (1997). Determination of DNA sequences required for regulated *Mycobacterium tuberculosis RecA* expression in response to DNA-damaging agents suggests that two modes of regulation exist. J. Bacteriol. Jun;179(11):3509-18.

Mulder, M.A., Zappe, H. and Steyn, L.M. (1999). The *Mycobacterium tuberculosis katG* promoter region contains a novel upstream activator. Microbiology, Sep;145 (Pt 9):2507-18.

UNIVERSITY of the
WESTERN CAPE

Mulder M.A., Zappe, H. and Steyn, L.M. (1997). Mycobacterial promoters. Tuber. Lung Dis;78(5-6):211-23.

Murphy, S., Di Liegro, C. and Melli, M. (1987). The in vitro transcription of the 7SK RNA gene by RNA polymerase III is dependent only on the presence of an upstream promoter. Cell. Oct. 9;51(1):81-7.

Murray, A., Winter, N., Lagranderie, M., Hill, D.F., Rauzie, J., Timm, J., Leclerc, C., Moriarty, K.M., Gheorghiu, M. and Gicquel, B. (1992). Expression of *Escherichia coli* beta-galactosidase in Mycobacterium bovis BCG using an expression system isolated from *Mycobacterium paratuberculosis* which induced humoral and cellular immune responses. Mol. Microbiol. Nov;6(22):3331-42.

Mustafa, A.S., Gill, H.K., Nerland, A., Britton, W.J., Mehra, V., Bloom, B.R., Young, R.A. and Godal, T. (1987). Human T-cell clones recognize a major *M.leprae* protein antigen expressed in *E. coli*. Nature Jan 2-8;319(6048):63-6.

Musters, W., Knol, J., Maas, P., Dekker, A.F., van Heerikhuizen, H. and Planta, R.J. (1989). Linker scanning of the yeast RNA polymerase I promoter. Nucleic Acids Res. Dec 11;17(23):9661-78.

Nakayama, M., Fujita, N., Ohama, T., Osawa, S. and Ishihama, A. (1989). Micrococcus luteus, a bacterium with a high genomic G+C content, contains *Escherichia coli*-type promoters. Mol. Gen. Genet. Sep;218(3):384-9.

Neuwald, A.F., Liu, J.S. Lipman, D.J. and Lawrence, C.E. (1997). Extracting protein alignment models from the sequence database. Nucleic Acids Res. May 1;25(9):1665-77.

UNIVERSITY of the
WESTERN CAPE

Newlands, J.T., Josaitis, C.A., Ross, W. and Gourse, R.L. (1992). Both fis-dependent and factor-independent upstream activation of the *rrnB* P1 promoter are face of the helix dependent. Nucleic Acids Res. Feb. 25;20(4):719-26.

Nikolov, D.B. and Burley, S.K. (1997). RNA polymerase II transcription initiation: a structural view. Proc. Natl. Acad. Sci. U S A. Jan. 7;94(1):15-22.

O'Neill M.C. (1989). Consensus methods for finding and ranking DNA binding sites. Application to *Escherichia coli* promoters. J. Mol. Biol. May 20;207(2):301-10.

O'Neill, M.C. (1991). Training back-propagation neural networks to define and detect DNA- binding sites. Nucleic Acids Res. 19(2):313-8.

O'Neill, M.C. (1992). *Escherichia coli* promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. Nucleic Acids Res. 20(13):3471-7.

Oppon, J. and Hide, W. (1998). A statistical model for Prokaryotic Promoter Prediction. Genome. Informatics '98. 271-273.

UNIVERSITY of the
WESTERN CAPE

O'Sullivan, A and Sueoka, N. (1967). Sequential replication of the *Bacillus subtilis* chromosome. IV. Genetic mapping by density transfer experiment. J. Mol. Biol. Jul. 28;27(2):349-68.

Ozoline, O.N. and Tsyganov, M.A. (1995). Structure of open promoter complexes with *Escherichia coli* RNA polymerase as revealed by the DNase I footprinting technique: compilation analysis. Nucleic Acids Res. Nov 25;23(22):4533-41.

Ozoline, O.N., Deev, A.A. and Arkhipova, M.V. (1997). Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. Nucleic Acids Res. 23:4703-9.

Pedersen, A.G. and Engelbrecht, J. (1995). Investigations of *Escherichia coli* promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. Ismb. 3:292-9.
Pedersen, A.G., Baldi, P., Brunak, S. and Chauvin, Y. (1996). Characterization of prokaryotic and eukaryotion promoters using hidden Markov models. Ismb. 4:182-91.

Pape, L.K., Windle, J.J. and Sollner-Webb, B. (1990). Half helical turn spacing changes convert a frog into a mouse rDNA promoter: a distant upstream domain determines the helix face of the initiation site. Genes Dev. Jan;4(1):52-62.

Paule, M.R. and White, R.J. (2000). Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res. Mar 15;28(6):1283-98.

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M. and Saccone, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res. 20(11):2871-5.

Pikaard, C.S., McStay, B., Schultz, M.C., Bell, S.P., Reeder, R.H. (1989). The Xenopus ribosomal gene enhancers bind an essential polymerase I transcription factor *xUBF*. Genes Dev. Nov;3(11):1779-88.

Pikaard, C.S., Smith, S.D., Reeder, R.H. and Rothblum, L. (1990). *rUBF*, an RNA polymerase I transcription factor from rats, produces DNase I footprints identical to those produced by *xUBF*, its homolog from frogs. Mol. Cell Biol. Jul;10(7):3810-2.

Prestridge, D.S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. J. Mol. Biol. Jun. 23;249(5):923-32.

Ptitsyn, A.A., Rogozin, I.B., Grigorovich, D.A., Strelets, V.B., Kel, A.E., Milanezi, L. and Kolchanov, N.A. (1996). Computer system "AutoGene" for automatic analysis of nucleotide sequences. Mol Biol (Mosk) Mar-Apr;30(2):432-41.

UNIVERSITY of the WESTERN CAPE

Ramesh, G. and Gopinathan, K.P. (1995). Cloning and characterization of mycobacteriophage I3 promoters. Indian J. Biochem. Biophys. Dec;32(6):361-7.

Read, C., Larose, A.M., Leblanc, B., Bannister, A.J., Firek, S., Smith, D.R. and Moss, T. (1992). High resolution studies of the Xenopus laevis ribosomal gene promoter in vivo and in vitro. J Biol. Chem. Jun 5;267(16):10961-7.

Reeder, R.H. (1984). Enhancers and ribosomal gene spacers. Cell Sep;38(2):349-51.

Reeder, R.H. (1990). rRNA synthesis in the nucleolus. Trends Genet. Dec;6(12):390-5.

Roeder, G.S., Voelkel-Meiman, K. and Keil, R.L. (1987). Recombination-stimulating sequences in yeast ribosomal DNA correspond to sequences regulating transcription by RNA polymerase I. Cell Mar 27;48(6):1071-9.

Rosenberg, M. and Court, D. (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. 13:319-53.

Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C, Ishihama, A., Severinov. K. and Gourse, R.L. (1993). A third recognition element in bacterial promoters DNA binding by the alpha subunit of RNA polymerase. Science 262(5138):1407-13.

Shmatkov, A.M., Melikyan, A.A., Chernousko, F.L., Borodovsky, M. (1999). Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. Bioinformatics. Nov;15(11):874-86.

Shulman, M.J., Steinberg, C.M. and Westmoreland, N. (1981). The coding function of nucleotide sequences can be discerned by statistical analysis. J. Theor. Biol. Feb 7;88(3):409- 20.

WESTERN CAPE

Siebenlist, U., Simpson, R.B. and Gilbert, W. (1980). *E.coli* RNA polymerase interacts homologously with two different promoters. Cell Jun;20(2):269-81.

Smale, S.T. and Baltimore, D. (1989). The "initiator" as a transcription control element. Cell Apr 7;57(1):103-13.

Sharp, S.J. and Garcia, A.D. (1988). Transcription of the Drosophila melanogaster 5S RNA gene requires an upstream promoter and four intragenic sequence elements. Mol. Cell Biol Mar;8(3):1266-74.

Sherman, D.R., Mdluli, K., Hickey, M.J., Arain, T.M., Morris, S.L., Barry, C.E. and Stover, C.K. (1996). Compensatory ahpC gene expression in isoniazid-resistant *Mycobacterium tuberculosis*. Science Jun.14;272(5268):1641-3.

Slonim, D., Kruglyak, L., Stein, L. and Lander, E. (1997). Building human genome maps with radiation hybrids. J. Comput. Biol. Winter;4(4):487-504.

Smith, S.D., O'Mahony, D.J., Kinsella, B.T. and Rothblum, L.I. (1993). Transcription from the rat 45S ribosomal DNA promoter does not require the factor UBF. Gene Expr. 3(3):229-36.

Sollner-Webb, B. and Mougey, E.B. (1991). News from the nucleolus: rRNA gene expression. Trends Biochem. Sci. Feb;16(2):58-62.

Sollner-Webb, B. and Tower J. (1986) Transcription of cloned eukaryotic ribosomal RNA genes. Annu Rev Biochem 55:801-30.

Sollner-Webb, B., Tower, J. and Culotta, V.C. (1986) Factors and nucleotide sequences that direct ribosomal DNA transcription and their relationship to the stable transcription complex. Mol. Cell Biol. Oct;6(10):3451-62.

Solovyev, V. and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Ismb. 5:294-302.

Sommervile, J. (1984) RNA polymerase I promoters and transcription factors. Nature. Jul. 19- 25;310(5974):189-90.

Sonnhammer, E.L., Eddy, S.R, Birney, E., Bateman, A. and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. Jan 1;26(1):320-2.

Spiegelhalter, F. and Bremer, E. (1996). Osmoregulation of the *opuE* proline transport gene from *Bacillus subtilis*: contributions of the sigma A- and sigma B-dependent stress-responsive promoters. Mol. Microbiol. Jul;29(1):285-96.

Staden, R. and McLachlan, A.D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Res. Jan 11;10(1):141-56.

Stultz, C.M., White, J.V. and Smith, T.F. (1993). Structural analysis based on state-space modeling. Protein Sci. Mar;2(3):305-14.

Suzuki, Y., Nagata, A. and Yamada, T. (1991). Analysis of the promoter region in the rRNA operon from *Mycobacterium* bovis BCG. Antonie Van Leeuwenhoek Jul;60(1):7-11.

Sunyaev, S.R., Eisenhaber, F. Argos, P., Kuznetsov, E.N. and Tumanyan, V.G. (1998). Are knowledge-based potentials derived from protein structure sets discriminative with respect to amino acid types? Proteins. May 15;31(3):225-46.

UNIVERSITY of the
WESTERN CAPE

Szoke, P., Allen T.L. and deHaseth, P.L. (1987). Promoter recognition by *Escherichia coli* RNA polymerase: effects of base substitutions in the -10 and -35 regions. Biochemistry Sep. 22;26(19):6188-94.

Taylor, W.R. (1996). A non-local gap-penalty for profile alignment. Bull Math Biol. Jan;58(1):1-18.

Theeragool, G., Miyao, A., Yamada, K., Sato, T., Kobayashi, Y. (1993). In vivo expression of the *Bacillus subtilis spoVE* gene. J. Bacteriol. Jul;175(13):4071-80.

Thole, J.E., Dauwerse, H.G., Das P.K., Groothuis, D.G., Schouls L.M. and van Embden, J.D. (1985). Cloning of *Mycobacterium bovis* BCG DNA and expression of antigens in *Escherichia coli*. Infect. Immun. Dec;50(3):800-6.

Thole, J.E., Keulen, W.J., De Bruyn, J., Kolk, A.H., Groothuis, D.G., Berwald, L.G., Tiesjema, R.H. van Embden, J.D. (1987). Characterization, sequence determination, and immunogenicity of a 64-kilodalton protein of *Mycobacterium bovis* BCG expressed in *Escherichia coli* K-12 Infect Immun. Jun;55(6):1466-75.

Thomas, T.J, Andrews, R.E. Jr. and Thoen, C.O. (1992). Molecular cloning and characterization of Mycobacterium paratuberculosis promoters in *Escherichia coli*. Vet. Microbiol Oct;32(3-4):351-62.

Thorne, J.L. and Jones, D.T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. Oct 25;263(2):196-208.

Tichelaar, J.W., Knerer, B., Vrabel, A. and Wieben, E.D. (1994). Transcription of a variant human U6 small nuclear RNA gene is controlled by a novel, internal RNA polymerase III promoter. Mol. Cell Biol. Aug;14(8):5450-7.

WESTERN CAPE

Timm, J, Perilli, M.G., Duez, C., Trias, J., Orefici, G., Fattorini, L., Amicosante, G., Oratore, A., Joris, B. and Frere, J.M. (1994). Transcription and expression analysis, using lacZ and phoA gene fusions, of *Mycobacterium fortuitum* beta-lactamase genes cloned from a natural isolate and a high-level beta-lactamase producer. Mol. Microbiol. May;12(3):491-504.

Tjian. R. and Maniatis, T. (1994). Transcriptional activation: a complex puzzle with few easy pieces. Cell Apr 8;77(1):5-8.

Trifonov, E.N. and Sussman, J.L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc. Natl. Acad. Sci. U.S.A. Jul;77(7):3816-20.

Van den Berg, E.A., Geerse, R.H., Pannekoek, H. and van de Putte, P. (1983). In vivo transcription of the *E.coli uvrB* gene: both promoters are inducible by UV. Both promoters are

inducible by UV. Nucleic Acids Res. Jul. 11;11(13):4355-63.

Verma, A., Kinger, A.K. and Tyagi, J.S. (1994). Functional analysis of transcription of the *Mycobacterium tuberculosis* 16S rDNA-encoding gene. Gene Oct 11;148(1):113-8.

Vold B.S., Okamoto K, Murphy, B.J., Green C.J. (1988). Transcriptional analysis of *Bacillus subtilis* rRNA-tRNA operons. The tRNA gene cluster of rrnB has an internal promoter. J. Biol. Chem. Oct 5;263(28):14480-4.

Weis, L. and Reinberg, D. (1992). Transcription by RNA polymerase II:initiator-directed formation of transcription-competent complexes. FASEB J Nov;6(14):3300-9.

UNIVERSITY of the
WESTERN CAPE

Yada, T, Ishikawa, M., Tanaka, H., and Asai, K. (1996). Extraction of hidden Markov model representations of signal patterns in DNA sequences. Pac. Symp. Biocomput:686-96.

Yada, T., Ishikawa, M., Tanaka, H. and Shinnick, T.M. (1987). The 65-kilodalton antigen of *Mycobacterium tuberculosis*. J. Bacteriol. Mar;169(3):1080-8.

Yamada, M., Kubo, M., Miyake, T., Sakaguchi, R., Higo, Y. and Imanaka, T. (1991). Promoter sequence analysis in *Bacillus* and *Escherichia*: construction of strong promoters in *E.coli*. Gene., Mar. 1;99(1):109-14.

Yuan, Y. and Reddy, R. (1991). 5' flanking sequences of human MRP/7-2 RNA gene are required and sufficient for the transcription by RNA polymerase III. Biochim. Biophys. Acta. May 2;1089(1):33-9.

Wilson, T.M. and Collins, D.M. (1996). *ahpC*, a gene involved in isoniazid resistance of the *Mycobacterium tuberculosis* complex. Mol. Microbiol. Mar;19(5):1025-34.

Wang, Y. and Stumph, W.E. (1995). RNA polymerase II/III transcription specificity determined by TATA box orientation. Proc. Natl. Acad. Sci. (U S A) Sep 12;92(19):8606-10.

UNIVERSITY of the
WESTERN CAPE