CHAPTER 2

MATERIALS AND METHODS

2.1 COMPARATIVE GENOMICS

The genomes to be compared were chosen from enterobacteriaceae. The non pathogenic genome of *Escherichia coli* K12 was used as a control. The pathogenic genomes analyzed were that of *Salmonella typhi* CT18 (Parkhill et al 2001), *Salmonella typhimurium* (McClelland et al 2001), *Salmonella typhi* Ty2, *Salmonella paratyphi*, *Escherichia coli* MG1655 K12 (Blattner et al 1997), *Escherichia coli* O157:H7-EDL933, *Escherichia coli* O157:H7 (Perna et al 2001), *Escherichia coli* CFT073. The protein sequence files (.faa files) and the respective annotation files were obtained from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Annotations for ORF start sites were used to extract ORF upstream region from genomes.

BLAST (Altschul et al 1997) was used with an e-value cutoff of 1E-4. Bidirectional best hit method was used to identify orthologs. The common minimal set of genes in *Salmonellae* (pathogenic genomes) was compared against the same from the genus *Escherichia*. The resultant proteins were presented as the common minimal pathogenesis set in enterobacteriaceae genomes. The proteins encoded by plasmids were not taken into account. Standalone BLAST was obtained from the NCBI ftp server.
2.2 OLIGONUCLEOTIDE PROFILING

Word usage in gene proximal upstream regions has been documented. To count the word frequencies ‘compseq’ program in EMBOSS (Rice et al., 2000) was used. Words were counted in 4 contiguous non-overlapping fragments (100 bases per fragment) of DNA for each ORF predicted. The fragments spanned the -100, -200, -300 and -400 regions. Words size was set to six. Control sets were created from randomly chosen sequences of the genome. Three such control sets were created. All sequence sets had the same number of sequences. The occurrence of each hexanucleotide in test sets (F1, F2, F3 and F4) was normalized with their respective average occurrence in the random sets. Based on the normalized score, the words were classified as enriched, random or rare. A hexanucleotide was termed enriched in a particular fragment set if its occurrence is at least twice the random occurrence. Similarly a word with half or less of its expected frequency was termed rare. The statistical significance of these cut-offs are discussed in the discussion section.

2.2.1 Word co-occurrence analysis

Co-occurrence of small words (di- and trinucleotides) with a spacer (d=4 to 30 bases) were counted in 5 contiguous overlapping sequence sets spanning -1 to -300 bases upstream of all the ORFs in the *Escherichia coli* genome. The fragments spanned -1 to -100, -51 to -150, -101 to -200, -151 to -250 and -201 to -300. Approximately 1.1x10^5 combinations were searched and the results analyzed. A similar search was performed on three equal sized random sequence set. The average of co-occurrences in the random sets was taken to be the random occurrence set. The co-occurrence in the 5 fragment sets were compared against the random set. Combinations were chosen that had at least two fold changes in occurrences in the -100 region.
2.3 PROMOTER PREDICTION

2.3.1 Extraction of Sequences

Oligonucleotide profiling was performed on 140 available bacterial genomes. Of these, in 106 genomes, the hexanucleotide TATAAT was enriched in the -100 region of ORFs. From all these genomes, all occurrences of TATAAT between the -50 and -20 were extracted along with 50 bases upstream. Thus for each occurrence of TATAAT, a 56 base sequence was extracted with the core signal at its 3' end. From 106 genomes, ~11100 sequences were extracted.

2.3.2 Creation and training of the matrix

The information in all these sequences was collated into a probabilistic profile. This profile was used to score each of these sequences. The scoring scheme was additive, with the score of each sequence being the sum of position specific probabilities of all its individual bases. The average score of a sequence in this set was calculated and all the sequences that scored above average were used for creating the next scoring matrix. This process was iterated until secondary signals were seen in the matrix.

In DNA, there are four possible character states. In a random sequence,

\[ p(N) = p(A) = p(C) = p(G) = p(T) = 0.25; \]

In the matrix, we set a stringent threshold of 0.6 or more for a significant secondary signal (SSS). After 4 iterations of pruning the sequence set the matrix exhibited one SSS at the extended -10 region with \( p(T) = 0.64 \). The \( p(G) \) of the associated G was the highest in the profile, with a value of
more than 0.4. This value of $p(G)$ is much higher than the average which is ~0.2. This matrix was then used for searching the gene upstream regions for promoters.

### 2.3.3 Selection of Sequences

Experimentally verified promoters of *Escherichia coli* were obtained from RegulonDB (Salgado et al., 2004). There were a total of 663 $\sigma$70 promoters for 528 genes. Promoter sequences were pruned to 56 base stretches with the 3’ end being the Pribnow box. The 663 promoters are given in supplementary information 3. The method was capable of predicting only one promoter per sequence. For prediction, the upstream region (till -300) of all 4311 genes was taken. These sequences were scored using the scoring matrix. The scoring scheme was an additive one, with the score of a sequence fragment equaling that of the sum of position specific probabilities of its character states. The highest scoring segment in each 300 base region was reported as the promoter. Positive hits were ascertained with the prediction results for the 528 genes whose promoters are known using BLASTN.

To create a negative control, we chose TATAAT occurrences in the *Escherichia coli* K12 genome that are at least 300 bases away from the nearest ORF. Sequences were extracted from these loci as described before with the TATAAT tail. Of these sequences containing tetra-A and tetra-T were removed. A profile was created from these sequences and used for scoring sequences.

### 2.3.4 Specificity, Sensitivity, and Accuracy tests

After ascertaining the prediction capacity, we assessed the sensitivity of our method. For this, we calculated the maximum error ($E_{\text{max}}$).
in number of bases, present in our prediction. Then, setting an error approximation of $E_{\text{max}}$ on both sides of the prediction, we were able to calculate the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Thus the total size of the DNA fragment analyzed was given by

$$S_{\text{Tot}} = S_{\text{Pred}} + (2 \times E_{\text{max}}).$$

Where, $S_{\text{Pred}}$ in this case was 56 bases and $E_{\text{max}}$ was 29 bases. Thus the total number of bases taken into analysis was 114 bases. For each of the predictions, the number of TP, FP, FN and TN were calculated and the sensitivity expressed as percentages of perfect matches. A perfect match would have

$$TP = 56; \; TN = 58; \; FP = 0; \; FN = 0.$$  

The specificity, sensitivity and accuracy, as defined by Huerta and Collado-Vides (Huerta and Collado-Vides, 2003) are given by

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$  
$$\text{Specificity} = \frac{TN}{TN + FP}$$  
$$\text{Accuracy} = \frac{TP}{TP + FP}$$

The results of nucleotide level sensitivities and specificities are discussed in results section. The method for prediction is illustrated in Figures 2.1 and 2.2 illustrates the test for specificity.
Figure 2.1 Method for promoter prediction

Sequences containing TATAAT in the region of interest are extracted and collated into a profile. The profile is then trained over itself until secondary signals are found. Once secondary signals are identified, training is stopped and resulting matrix is used to predict promoters.

Figure 2.2 Illustration for sensitivity tests in promoter prediction
Illustration for assessing sensitivity, specificity and accuracy for the predictions. Each base will be given one of the 4 flags TP, FP, FN or TN. These values are then used to calculate abovesaid parameters. The formulae are discussed in methods section.