2158–2165 Nucleic Acids Research, 1994, Vol. 22, No. 11

Analysis of *E.coli* promoter structures using neural networks

Indu Mahadevan and Indira Ghosh*
Astra Research Centre India, 18th Cross Road, Malleswaram, Bangalore 560 003, India

Received August 24, 1993; Revised January 6, 1994; Accepted April 7, 1994

ABSTRACT

Backpropagation neural network is trained to identify *E.coli* promoters of all spacing classes (15 to 21). A three module approach is employed wherein the first neural net module predicts the consensus boxes, the second module aligns the promoters to a length of 65 bases and the third neural net module predicts the entire sequence of 65 bases taking care of the possible interdependencies between the bases in the promoters. The networks were trained with 106 promoters and random sequences which were 60% AT rich and tested on 126 promoters (Bacterial, Mutant and Phage promoters). The network was 98% successful in promoter recognition and 90.2% successful in non-promoter recognition when tested on 5000 randomly generated sequences. The network was further trained with 11 mutated non-promoters and 8 mutated promoters of the p22ant promoter. The testing set with 7 mutated promoters and 13 mutated non-promoters of p22ant were identified.The network was upgraded using total 1665 data of promoters and non-promoters to identify any promoter sequences in the gene sequences.The network identified the locations of P1, P2 and P3 promoters in the pBR322 plasmid. A search for the start codon, Ribosomal Binding Site and the stop codon by a string search procedure has also been added to find the possible promoters that can yield protein products. The network was also successfully tested on a synthetic plasmid pWM528.

INTRODUCTION

Prediction of promoter sequence in prokaryotes using statistical and neural network approaches have shown that the simple consensus sequence prediction is not sufficient to identify putative promoters. Application of neural networks for recognition of promoters is quite evident by the number of publications in the last few years. A neural net that is 94 to 97% successful in promoter recognition and 96 to 98% successful in classifying an arbitrary sequence has been claimed (1). In this network, the relation between the −10 and −35 boxes and the possible influence of the spacer and other regions on the sequence being sensed by RNA polymerase as a promoter was not considered since the network learnt only the boxes. This network may not be successful in identification of a sequence change from a promoter to a non- promoter caused by single-point mutations in the spacers. Another study (2) used a three stage network wherein the first and second stages were trained with the extended −10 and −35 boxes. The third stage was trained with an aligned sequence of −10 and −35 boxes. The prediction accuracy was apparently 98% for promoters and 98.45% for non-promoters. The training set included 80 promoters and the network was tested on 30 promoters and 1500 non-promoters which included only the 16, 17 and 18 spacer classes only. On the other hand a neural net for the prediction of only promoters with 17mer spacers reported an overall prediction accuracy of 80% (3). In a subsequent paper separate networks were used for each class of 16, 17 and 18 length spacers (4). It was reported that a network which was tried for identifying the promoters of all the spacer lengths resulted in a promoter recognition of 60% only. All the above methods attempted to predict promoter sequences in the 5′ upstream region of the coding region of the gene giving maximum weightage to the consensus sequences. Earlier studies have shown that the consensus nature in these boxes are preserved highly for spacer class 16 promoters and consensus sequence changes with the spacer length between them (5,6). It has been stated (7),using expectation maximization algorithm that, as many as eight positions in the spacer region may contribute to promoter specificity. Hence, there are attempts to integrate these characteristics of the promoters by using a series of networks (4), but the method of generating promoters by shuffling the sequence might have subdued the effect of positional dependency, i.e. the dependency of a base in a particular position may have had on other bases at other positions. Also, since a sequence is passed through a series of networks and a polling is done of the various outputs obtained, we dont really know if a sequence is a promoter or not if different networks give contradicting results.

We have suggested an integrated approach, wherein the consensus boxes are searched by a backpropagation network, alignment is done to retain the information in the spacer and other regions and finally the aligned sequence of length 65 bases is used to train another network to identify promoter sequences of varying spacer length. The first stage of the program recognizes the RNA polymerase binding site, the −35 and −10 regions. The second stage aligns the sequences introducing blanks in such

*To whom correspondence should be addressed

a way that irrespective of the spacer length (bases between the consensus) the distance between the RNA polymerase binding regions is retained. The third and final stage learns the promoters which are aligned to take into account the varying spacer lengths. This module also includes in its computation any possible interdependence of bases in various positions with respect to promoter recognition. The highlight of this method is that a combination of two neural networks and an 'align program' are used to identify promoters of all spacer lengths instead of using a separate network for each promoter spacer length. The network was initially, trained with a set 106 promoters and was tested on another set of 126 promoters (5,8,9) as well as 5000 randomly generated sequences. The network was successful 98% in recognizing promoters and 90.2% in recognizing non-promoters. The network was used in the identification of single-point mutations in P22ant promoter and the promoter sites in the pBR322 plasmid also.

DETAILS OF THE NETWORK

The Backpropagation network has been developed on a UNIX-based workstation using the C language. The networks used for recognition are three-layered feed-forward networks (10). The training of the network is done in two stages called the forward and reverse passes. In the forward pass outputs are calculated by summing the products of all inputs to a neuron with their respective weights, adding a bias value to it and operating the sum on a sigmoid function F. In the reverse pass error values are backpropagated and the weight and bias values are changed.

The output of neuron j is given by

$$Output_j = F\ (Input_j) = 1/\ (1 + exp^{(-Input_j)})$$

where

$$Input_j = \Sigma_i\ (\text{Output}_i * \text{weight}_{ji}) + \text{bias}_j$$

$Output_i$ is the output of unit i in the previous layer, $weight_{ji}$ is the weight connections between units i and j and $bias_j$ is the threshold value. Training consists of presenting each training set pattern at the input units and iteratively minimizing the difference between the output of the network and desired target value.

Each base is represented by a four bit pattern. The bit patterns used are A=0001, T=0010, G=0100 and C=1000. The network learns a 1.0 if the sequence is a promoter and a 0.0 otherwise.

The method used for the classification process consisted of three modules

1. Module I had two neural networks which learnt only the −10 and −35 boxes. Given any sequence, the network checked if there were −10 and −35 boxes with 15 to 21 bases separating them. The starting point had to be specified or else, the network considered any A or G base within the first 10 bases from 3′ end towards −10 box as the starting point.
2. After the module I had decided upon the two boxes, the module II aligned the sequence depending on the spacer length.
3. Module III learnt and predicted the aligned sequences.

MODULE I OF THE NETWORK

This module had two neural networks which learnt the −35 and −10 boxes separately. The input layer had 24 neurons (6×4 neurons). The output layer had 1 neuron and the hidden layer had 2 neurons. The network learnt a 1.0 for possible promoter boxes and a 0.0 for possible non-promoter boxes.

The learning set consisted of 106 promoters. The duplicate boxes obtained from any promoters were eliminated. The non-promoters used were random hexamers which matched with not more than two bases with the conserved −10 and −35 boxes. After removing duplicates, we had 58 unique boxes of the −10 and 72 unique boxes of the −35 regions to be taught to the network. These boxes were added with non-promoters in the ratio of 1:1 (promoter:non-promoter) for the learning proess.

MODULE II OF THE NETWORK

This program aligned the sequences with respect to the boxes and spacers identified by module I of the network. Boxes with output greater than 0.8 and having a spacer of 15 to 21 bases between them were considered as potential promoters and aligned. A cut-off value of 0.8 was chosen for promoter recognition in module I. The sequences were aligned by inserting the required blanks according to the method mentioned in the subsection below. The aligned sequence output of this network was used as input to the network in module III.

Alignment of promoters

The alignment of the promoters in our program for the module III of the network which learnt the entire sequence was done as follows

1. 14 bases on the 5′ end of -35 box were considered. If all the 14 bases were not available (in ref 5, 8, and 9), blanks (−) were introduced into the sequence. Blanks (−) meant that the network did not learn at the positions where the blanks occured when it was learning a particular pattern. This allows inclusion of weakly conserved A at position −45.
2. −35 box was extended on the 3′ side to include 5 bases from the spacer which took care of the weakly conserved T in the spacer.
3. The rest of the spacer had blanks inserted in such a way that the entire spacer length was 21. The blanks were introduced at the 5′ end of the spacer sequence. Introduction of the blanks at the 5′ end helped in aligning the weakly conserved T at around −18 position and weakly conserved TG at around −15 and −16 positions (11).
4. −10 box was followed by three bases on 3′ end.
5. Other bases and blanks were introduced such that the +1 transcription initiation point was 12 bases away from the last T of the −10 box towards the 3′ end including the +1 itself. (This was done so because a few of the promoters given in (5) had +1 points identified as far as 12 bases (maximum) from the last T of the −10 box).
6. 6 bases followed the +1 point in the 3′ side. Blanks (−) were introduced if bases in these positions were not available. The length of the aligned sequence would be then 65 bases. All the promoters used in the study are shown in Table 1. An example of alignment with promoter araE is shown in Fig 1.

MODULE III OF THE NETWORK

This network learnt the aligned 65 length sequences of 106 promoters. The network had 260 neurons (65×4) in the input layer. The output layer had 1 neuron. The number of neurons in the hidden layer was varied between 2 and 12. The predictability of the network did not vary very much. So 7 neurons were used in the hidden layer for the training process. The network was also tried to train without a hidden layer which

```
                -35          Spacer          -10
CTGTTTCCGAC CTGACA CCTGCGTGAGTTGTTCACG TATTTT TTCACTATgTCTTAC

↓     14      ↓ -35  ‖ 5 ‖▲Rest of Spacer▲ -10  |3|         • 6  •
---CTGTTTCCGACCTGACACCTGC--GTGAGTTGTTCACGTATTTTTTC---ACTATgTCTTAC
```

Figure 1. The aligned sequence of araE is shown. The regions between the various pairs of markers denote the number of bases. 14 bases between 5′ end and −35 box are represented between ↓. The 5 bases retained after −35 box towards 3′ end are highlighted with ||. The rest of the spacer which is aligned with blanks is shown between Δ. The 3 bases following the −35 box are shown within |. The 6 bases after the +1 point are represented within •. The base in lower case represents the +1 point.

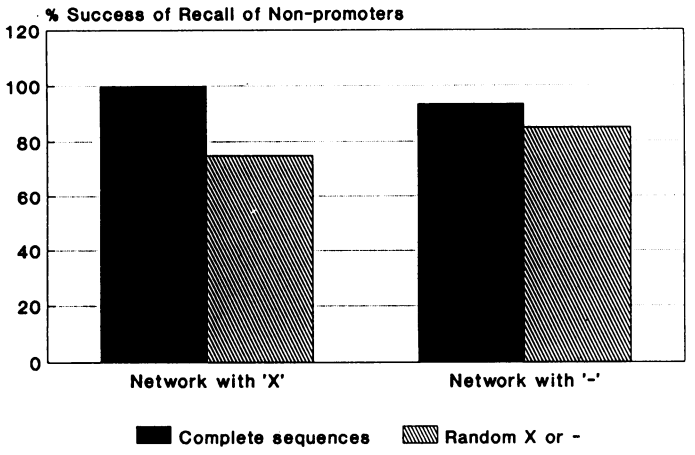

Figure 2. The effect of using '*X*' and '−' during non-promoter recognition is shown. Left side represents the percentage success of recall of network which has learnt promoters with '*X*' and non- promoters without '*X*' and which are tested on non-promoters with and without '*X*'. Right side represents the percentage success of recall of a network that has learnt promoters with blank (−) and non-promoters without '−' and tested on non-promoters with and without '−'. '*X*' represents a pattern (1111) and '−' means no learning. Filled box represents complete sequences and hashed box represents sequences with random '*X*' or '−' introduced. (See text)

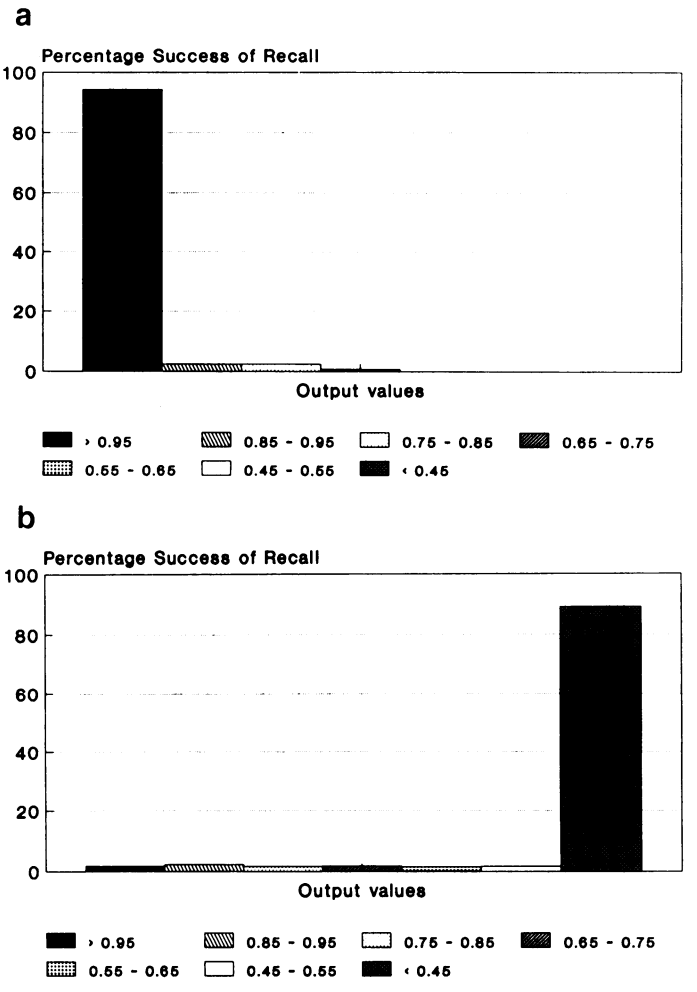

Figure 3. a shows the percentage of promoters at various output value ranges of the 126 promoters tested. **b.** shows the percentage of non-promoters at various output value ranges of the random 5000 sequences tested.

would reduce the number of free parameters drastically. But it was unsuccessful with regard to percentage of recall in both the cases, promoters and non-promoters. This emphasizes that the correlations in the positions of the promoter sequences are important.

A ratio of 1:2 of promoters to non-promoters was used in the training of the network. The choice of non-promoters for the learning process is discussed in the subsection below.

Choice of non-promoters

For our application, we chose sequences from the coding region of the gene which were more than 60% AT rich as non-promoters. There is a probability that these stretches might have sequences that resemble the consensus boxes in a promoter. Therefore, to make sure that the sequence considered did not have the required consensus sequence boxes of a promoter in the regions corresponding to the −10 and −35 regions, the homology methods used in (11) were considered. The TARGSEARCH search program discussed by the above authors was used to check if the sequence had a possibility of being a promoter or not. Since all promoters had homology scores greater than 30 (11), only sequences with homology scores below 15 were considered as non-promoters. Out of the 100 odd sequences obtained from the coding region of the bacterial *E.coli* gene, 30 sequences qualified as potential promoters. Since the learning set needed 106 non-promoters, these 70 odd non-promoters were scrambled to form another set of possible non-promoters which was again tested by the program TARGSEARCH. These set of sequences formed the non-promoters in the learning set.

Representing aligned promoters and non-promoters by '*X*' or '−'

When promoters were aligned, the positions with no bases, were represented by a blank (−). The choice of a blank over another pattern say '*X*' is discussed here. The spaces during alignment were initially represented by a pattern *X* (1111). During recall process we found that the '*X*'*s* were biasing the classification of the sequences towards the promoters. To find out the effect of learning a new pattern *X*, a blank (−) was represented using four neurons which were not triggered ie, the network did not learn that particular position of the pattern where a blank occured. Hence, that neuron did not perform any operation and did not contribute to the input of other neurons when this pattern is being learnt. A comparative test was done with the network learning

Table 1. Promoters used for training and testing the network

106 promoters used for training the network

MlRNA	ATGCGCAACGCGGGGTGACAAGGGC----GCGCAAACCCTCTATACTGCG----CGCCgAAGCTG
PA1	TATCAAAAAGAGTATTGACTTAAAG----TCTAACCTATAGGATACTTAC----AGCCaTCGAGA
PA2	ACGAAAAACAGGTATTGACAACATG---AAGTAACATGCAGTAAGATACA----AATCgCTAGGT
PA3	GTGAAACAAAACGGTTGACAACATG----AAGTAAACACGGTACGATGTA----CCACaTGAAAC
PG25	GAAAAATAAAATTCTTGATAAAATT----TTCCAATACTATTATAATATT-----GTTaTTAAAG
PH207	TTTAAAAAATTCATTTGCTAAACGC----TTCAAATTCTCGTATAATATA----CTTCaTAAATT
PJ5	TATAAAAACCGTTATTGACACAGGT----GGAAATTTAGAATATACTGTT-----AGTaAACCTA
PL	TATCTCTGGCGGTGTTGACATAAAT----ACCACTCGGCGGTATACTGAG-----CACaTCAGCA
PN25	CATAAAAAATTTATTTGCTTTCAGG----AAAATTTTTCTGTATAATAGA-----TTCaTAAATT
Pbla	TTTTTCTAAATACATTCAAATATGT----ATCCGCTCATGAGACAATAAC------CCtGATAAA
Pcon	ATTCACCGTCGTTGTTGACATTTTT----AAGCTTGGCGGTTATAATGGT-----ACCaTAAGGA
Pd/E20	ACTGCAAAAATAGTTTGACACCCTA----GCCGATAGGCTTTAAGATGTA-----CCCaGTTCGA
Plac	TAGGCACCCCAGGCTTTACACTTTA---TGCTTCCGGCTGGTATGTTGTG-----TGGaATTGTG
PlacUV5	TAGGCACCCCAGGCTTTACACTTTA---TGCTTCCGGCTGGTATAATGTG-----TGGaATTGTG
Poril	CTGTTGTTCAGTTTTTGAGTTGTGT----ATAACCCCTCATTCTGATCCC------AGcTTATAC
Porir	GATCGCACGATCTGTATACTTATTT---GAGTAAATTAACCCACGATCCC----AGCCaTTCTTC
PtacI	TCTGAAATGAGCTGTTGACAATTAA----TCATCGGCTCGTATAATGTG-----TGGaATTGTG
S10	TACTAGCAATACGCTTGCGTTCGGT-----GGTTAAGTATGTATAATGCG------CGgGCTTGT
alaS	AACGCATACGGTATTTTACCTTCCC---AGTCAAGAAAACTTATCTTATT-----CCCaCTTTTC
ampC	TGCTATCCTGACAGTTGTCACGCTG-----ATTGGTGTCGTTACAATCTA----ACGCaTCGCCA
araC	GCAAATAATCAATGTGGACTTTTCT----GCCGTGATTATAGACACTTTT---GTTACgCGTTTT
aroH	GTACTAGAGAACTAGTGCATTAGCT-----TATTTTTTTGTTATCATGCT----AACCaCCCGGC
bioA	GCCTTCTCCAAAACGTGTTTTTTGT---TGTTAATTCGGTGTAGACTTGT------AAaCCTAAA
bioB	TTGTCATAATCGACTTGTAAACCAA----ATTGAAAAGATTTAGGTTTAC---AAGTCtACACCG
deoP2	AATTGTGATGTGTATCGAAGTGTGT--TGCGGAGTAGATGTTAGAATACT-----AACaAACTCG
fol	CATCCTCGCACCAGTCGACGACGGT----TTACGCTTTACGTATAGTGGC-----GACaATTTTT
glnS	TAAAAAACTAACAGTTGTCAGCCTG----TCCCGCTTATAAGATCATACG------CCgTTATAC
his	ATATAAAAAAGTTCTTGCTTTCTAA---CGTGAAAGTGGTTTAGGTTAAA----AGACaTCAGTT
lacI	GACACCATCGAATGGCGCAAAACCT----TTCGCGGTATGGCATGATAGC----GCCCgGAAGAG
lacP1	TAGGCACCCCAGGCTTTACACTTTA---TGCTTCCGGCTCGTATGTTGTG-----TGGaATTGTG
leu	----------GTTGACATCCGT----TTTTGTATCCAGTAACTCTAA----AAGCaTATCGC
leuItr	TCGATAATTAACTATTGACGAAAAG-----CTGAAAACCACTAGAATGCG---CCTCCgTGGTAG
lexA	TGTGCAGTTTATGGTTCCAAAATCG----CCTTTTGCTGTATATACTCAC-----AGCaTAACTG
malEFG	AGGGGCAAGGAGGATGGAAAGAGGT-----TGCCGTATAAAGAAACTAGA----GTCCgTTTAGG
malK	CAGGGGGTGGAGGATTTAAGCCATC-----TCCTGATGACGCATAGTCAG-----CCCaTCATGA
recA	TTTCTACAAAACACTTGATACTGTA-----TGAGCATACAGTATAATTGC-----TTCaACAGAA
rplJ	TGTAAACTAATGCCTTTACGTGGGC----GGTGATTTTGTCTACAATTCT-----ACCcCCACGT
rpoA	TTCGCATATTTTTCTTGCAAAGTTG----GGTTGAGCTGGCTAGATTAGC------CAgCCAATC
rpoB	CGACTTAATATACTGCGACAGGACG-----TCCGTTCTGTGTAAATCGCA----ATGAaATGGTT
rrnABP1	TTTTAAATTTCCTCTTGTCAGGCCG-----GAATAACTCCCTATAATGCG---CCACCaCTGACA
rrnABP2	GCAAAAATAAATGCTTGACTCTGTA-----GCGGGAAGGCGTATTATGCA-----CACcCCGCGC
rrnDEX	CCTGAAATTCAGGGTTGACTCTGAA-----AGAGGAAAGCGTAATATACG----CCACcTCGCGA
rrnDP1	GATCAAAAAAATACTTGTGCAAAAA----ATTGGGATCCCTATAATGCG---CCTCCgTTGAGA
rrnEP1	CTGCAATTTTTCTATTGCGGCCTGC----GGAGAACTCCCTATAATGCG---CCTCCaTCGACA
rrnGP1	TTTATATTTTTCGCTTGTCAGGCCG-----GAATAACTCCCTATAATGCG---CCACCaCTGACA
rrnGP2	AAGCAAAGAAATGCTTGACTCTGTA-----GCGGGAAGGCGTATTATGCA-----CACcGCCGCG
rrnXP1	ATGCATTTTTCCGCTTGTCTTCCTG-----AGCCGACTCCCTATAATGCG---CCTCCaTCGACA
spc	CCGTTTATTTTTTCTACCCATATTC----TTGAAGCGGTGTTATAATGCC------GCgCCCTCG
str	TCGTTGTATATTTCTTGACACCTTT----TCGGCATCGCCCTAAAATTCG------GCgTCCTCA
supBE	CCTTGAAAAAGAGCTTGACGCTGCA----AGGCTCTATACGCATAATGCG----CCCCgCAACGC
thr	AAATTAAAATTTTATTGACTTAGGT----CACTAAATACTTTAACCAATA---TAGGCaTAGCGC
tnaA	AAACAATTTCAGAATAGACAAAAAC-----TCTGAGTGTAATAATGTAGC-----CTCgTGTCTT
trp	TCTGAAATGAGCTGTTGACAATTAA----TCATCGAACTAGTTAACTAGT----ACGCaAGTTCA
trpP2	ACCGGAAGAAAACCGTGACATTTTA----ACACGTTTGTTACAAGGTAAA-----GGCgACGCCG
trpR	TGGGGACGTCGTTACTGATCCGCAC---GTTTATGATATGCTATCGTACT----CTTTaGCGAGT
tufB	ATGCAATTTTTTACTTGCATGAACT----CGCATGTCTCCATAGAATGCG----CGCTaCTTGAT
uvrBP2	TCAGAAATATTATGGTGATGAACTG----TTTTTTTATCCAGTATAATTTG-----TTGgCATAAT
uvrBP3	ACAGTTATCCACTATTCCTGTGGAT----AACCATGTGTATTAGAGTTAG-----AAAaCACGAG
uvrbP1	TCCAGTATAATTTGTTGGCATAATT----AAGTACGACGAGTAAAATTAC------ATaCCTGCC
gyrB	-----CGGACGAAAATTCGAAGATGTTTACCGTGGAAAAGGGTAAAATAAC----GGATtAACCCA
rrnB P3	-CTATGATAAGGATTACTCATCTTA-TCCTTATCAAACCGTTAAAATGGG------CGgtgTGAG
araE	---CTGTTTCCGACCTGACACCTGC--GTGAGTTGTTCACGTATTTTTTC---ACTATgTCTTAC
glnL	CAATTCTCTGATGCTTCGCGCTTTT--TATCCGTAAAAAGCTATAATGCA------CTaAATGGT
tyrT/77	--ATTATTCTTTAATCGCCAGCAAA--AATAACTGGTTACCTTTAATCCG---TTACCgATGAAA
carAB P2	TAAGCAGATTTGCATTGATTTACGT---CATCATTGTGAATTAATATGCA-------AaTAAAGT
dapD	AAGTGCATCAGCGGTTGACAGAGGC---CCTCAATCCAAACGATAAAGGG------TGatgTGTT
dnaA 1p	TGCGGCGTAAATCGTGCCCGCCTCG---CGGCAGGATCGTTTACACTTAG-----CGAgTTCTGG
dnaK P1	TTTGCATCTCCCCCTTGATGACGTG---GTTTACGACCCCATTTAGTAGT-------caaCCGCA
gltA P2	AGTTGTTACAAACATTACCAGGAAA---AGCATATAATGCGTAAAAGTTA--------tGAAGTC
glyA/geneX	ACACCAAAGAACCATTTACATTGCA---GGGCTATTTTTTATAAGATGCA------TTtGAGATA
htpR P2	TTCACAAGCTTGCATTGAACTTGTG---GATAAAATCACGGTCTGATAAA-----ACAgTGAATG
pyrE P2	-GTAGGCGGTCATACTGCGGATCAT---AGACGTTCCTGTTTATAAAAGG------AGaGGTGG-
rpoD Pb	------AGCCAGGTCTGACCACCGG---GCAACTTTTAGAGCACTATCGT---GGTACaaaT---
ssb	TAGTAAAAGCGCTATTGGTAATGGT---ACAATCGCGCGTTTACACTTAT-----TCAgAACGAT
tonB	ATCGTCTTGCCTTATTGAATATGAT---TGCTATTTGCATTTAAAATCGA--------gACCTGG
trxA	CAGCTTACTATTGCTTTACGAAAGC---GTATCCGGTGAAATAAAGTCAA------CTaGTTGGT
tyrT/109	ACAGCGCGTCTTTGTTTACGGTAAT---CGAACGATTATTCTTTAATCGC------CAGcAAAAAT
tyrT/140	TTAAGTCGTCACTATACAAAGTACT---GGCACAGCGGGTCTTTGTTTAC--------gGTAATC
tyrT/6	--ATTTTTCTCAACGTAACACTTTA---CAGCGGCGCGTCATTTGATATG-----ATGcGCCCCG
motA	GCCCCAATCGCGCGTTAACGCCTGA---CGACTGAACATCCTGTCATGGT------CAaCAGTGG
sucAB	AAATGCAGGAAATCTTTAAAAACTG---CCCCTGACACTAAGACAGTTTT-----AAAaGGTTCC
aceEF	--ACGTAGACCTGTCTTATTGAGCT----TTCCGGCGAGAGTTCAATGGG--------aCAGGTC
ada	AAGATTGTTGGTTTTTGCGTGATGG----TGACCGGGCAGCCTAAAGGCT---------aTCCTTA
ampC/c16	--------GCTATCTTGACAGTTGT----CACGCTGATTGGTATCGTTAC---AATGTaACGTAT
argE P1	TTACGGCTGGTGGGTTTTTATTACGC----TCAACGTTAGTGTATTTTTAT------TCaTAAATA
argE P2	CCGCATCATTGCTTTGCGCTGAAAC----AGTCAAAGCGGTTATGTTCAT--------aTGCGGA
argF	ATTGTGAAATGGGGTTGCAAATGAA----TAATTACACATATAAAGTGAA----TTTTaATTCAA
argI	----------AGACTTGCAAATGAA----TAATCATCCATATAAATTGAA----TTTTaaTTCAT
argR	---TCGTCGCCGCGTTGCAGGAGCA----AGGCTTTGACAATATTAATCA----GTCTaaaGTCT
aroG	AGTGTAAAACCCCGTTTACACATTC----TGACGGAAGATATAGATTGGA-----AGTaTTGCAT
C62.5 P1	-CACCTGCTCTCGCTTGAAATTATT----CTCCCTTGTCCCCATCTCTCC-----CAcatCCTGT
carAB P1	ATCCCGCCATTAAGTTGACTTTTAG----CGCCCATATCTCCAGAATGCC-----GCCgTTTGCC
dnaQ P1	GCCAGGCGCTAAAGGTTTTCTCGCGT-----CCGCGATAGCGTAAAATAGC--------gccGTAA
frdABCD	---GATCTCGTCAAATTTCAGACTT-----ATCGATCAGACTATACTGTTGTACCTATaAAGG--
gltA P1	ATTCATTCGGGACAGTTATTAGTGG-----TAGACAAGTTTAATAATTCG------GAtTGCTAA
lep	TCCTCGCCTCAATGTTGTAGTGTAG-----AATGCGGCGTTTCTATTAAT------AcaGACGTT
cmpA	---GCCTGACGGAGTTCACACTTGT-----AAGTTTTCAACTACGTTGTA-----GACtTTAC--
pyrB1 P1	CTTTCACACTCCGCCCTATAAGTCG-----GATGAATGGAATAAAATGCA-------TatcTGAT
sdh P2	AGCTTCCGCGATTATGGGCAGCTTC-----TTCGTCAAATTTATCATGTG----GGGCaTCCTTA
tyrT/212	----------GATCATACCTAC-----ACAGCTGAAGATATGATGCG----CGCAgGTCGTG
uncI	TGGCTACTTATTGTTTGAAATCACG-----GGGGCCCACCGTATAATTTG-----ACCgCTTTTT
uvrD	TGGAAATTTCCCGCTTGGCATCTCT-----GACCTCGCTGATATAATCAG----CAAaTCTGTA
micF	GCGGAATGGCGAAATAAGCACCTAA-----CATCAAGCAATAATAATTCA----AGGTtAAAATC
cmpR	TTTCGCCGAATAAATTGTATACTTA------AGCTGCTGTTTAATATGCT------TTgTAACAA
rrnB P4	GCGTATCCGGTCACCTCTCACCTGA------CAGTTCGTGGTAAAATAGC-----CAAccTGTTC
tyrT/178	--TGCGCGCAGGTCGTGACGTCGAG------AAAAACGTCTTAAGTCGTG------CAcTATACA

60 bacterial promoters used for testing

Name	sequences	Score
MiPc 2	------GGAACACATTTAAAAACCC----TCCTAAGTTTTGTAATCTATA-----AAGttAGCAA	0.98
MiPe	TACCAAAAAGCACCTTTACATTAAG----CTTTTCAGTAATTATCTTTTT-----AGTaAGCTAG	0.99
araBAD	TTAGCGGATCCTACCTGACGCTTTT---TATCGCAACTCTCTACTGTTTC---TCCATaCCCGTT	0.99
argCBH	TTTGTTTTCATTGTTGACACACCT------CTGGTCATGATAGTATCAA---TATTCaTGCAGT	0.99
aroF	TACGAAAATATGGATTGAAAACTTT-----ACTTTATGTGTTATCGTTAC-----GTCaTCCTCG	0.99
cat	--ACGTTGATCGGCACGTAAGAGGT----TCCAACTTTCACCATAATGAA----ATAAgATCACT	0.99
cmpC	GTATCATATTCGTGTTGGATTATTC----TGCATTTTTGGGGAGAATGGA-------CtTGCCGA	0.99
cmpF	--------GGTAGGTAGCGAAACGT----TAGTTTGAATGGAAAGATGCC-----TGCaGACACA	0.97
crp	AAGCGAGACACCAGGAGACACAAAG----CGAAAGCTATGCTAAAACAGT-----CAGgATGCTA	0.98
cya	GTAGCGCATCTTTCTTTACGGTCAA----TCAGCAAGGTGTTAAATTGAT-----CACgTTTTAG	0.99
deo P3	ACACCAACTGTCTATCGCCGTATCA-----GCGAATAACGGTATACTGAT-----CTGaTCATTT	0.99
deoP1	CAGAAACGTTTTATTCGAACATCGA----TCTCGTCTTGTGTTAGAATTCT-----AACaTACGGT	0.99
divE	AAACAAATTAGGGGTTTACACGCCG----CATCGGGATGTTTATAGTGCG----CGTCaTTCCGG	0.99
dnaA 2p	TCTGTGAGAAACAGAAGATCTCTTG----CGCAGTTTAGGCTATGATCCG-------cggtccCG	0.99
dnaK P2	ATGAAATTGGGCAGTTGAAACCAGA-----CGTTTCGCCCCTATTACAGA-----CtcaCAACCA	0.99
fumA	GTACTAGTCTCAGTTTTTGTTAAAA----AAGTGTGTAGGATATTGTTAC---TCGCTtttAACA	0.85
galP1	CTAATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTAT-----TTCaTACCAT	0.97
galP2	CACTAATTTATTCCATGTCACACTT---TTCGCATCTTTGTTATGCTATG-----GTTaTTTCAT	0.98
glyA	TCCTTTGTCAAGACCTGTTATCGCA----CAATGATTCGGTTATACTGTT-------CgCCGTTG	0.99
gnd	GCATGGATAAGCTATTTATACTTTA----ATAAGTACTTTGTATACTTAT----TTGCgAACATT	0.99
groE	----TTTTTCCCCCTTGAAGGGGCG----AAGCCATCCCCATTTCTCTGG------TCaCCAGCC	0.96
hisA	GATCTACAAACTAATTAATAAATAG-TTAATTAACGCTCATCATTGTACA----ATGAaCTGTAC	0.81
hisBp	CCTCCAGTGCGGTTTAAATCTTT----GTGGGATCAGGGCATTATCTT--------acGTGAT	0.99
hisS	AAATAATAACGTGATGGGAAGCGCC----TCGCTTCCCGTGTATGATTGA----ACccgCATGGC	0.99
htpR P1	ACATTACGCCACTTACGCCTGAATA----ATAAAGCGTGTTATACTCTT-----TCCtGCAATG	0.99
htpR P3	AGCTTGCATTGAACTTGTGGATAAA----ATCACGGTCTGATAAAACAGT----GAATgATAACC	0.96
ilvGEDA	GGCCAAAAAATATCTTGTACTATTT---ACAAAACCTATGGTAACTCTTT----AGGCaTTCCTT	0.98
ilvIH P1	-CTCTGGCTGCCAATTGCTTAAGCA----AGATCGGACGGTTAATGTGTT------ttacacatt	0.99
ilvIH P2	GAGGATTTTATCGTTTCTTTTCACC----TTTCCTCCTGTTTATTCTTAT--------tACCCCG	0.99
ilvIH P3	-ATTTTAGGATTAATTAAAAAAATA----GAGAAATTGCTGTAAGTTGTG---GGATTcAGCCGA	0.90
ilvIH P4	TGTAGAATTTTATTCTGAATGTCTG----GGCTCTCTATTTTAGGATTAA----TTAAaAAAATA	0.99
lacP2	TTAATGTGAGTTAGCTCACTCATTA--GGCACCCCAGGCTTTACACTTTA-----TGCtTCCGGC	0.99
livJ	TGTCAAAATAGCTATTCCAATATCA----TAAAAATCGGGATATGTTTTA------GCaGAGTAT	0.99
lpd	---------TGTTGTTTAAAAATTG----TTAACAATTTTGTAAAATACC-GACGGAtagAACG-	0.99
lpp	CCATCAAAAAAATATTCTCAACATA---AAAAACTTTGTGTAATACTTGT-----AACgCTACAT	0.99
malPQ	ATCCCCGCAGGATGAGGAAGGTCAA----CATCGAGCCTGGCAAACTAGC-----GATaACGTTG	0.96
malT	AAAAACGTCATCGCTTGCATTAGAA----AGGTTTCTGGCCGACCTTATA----ACCaTTAATT	0.98
nanA	CGGCTCCAGGTTACTTCCCGTAGGA----TTCTTGCTTTAATAGTGGGAT----TAATttCCACA	0.79
netA P1	TTCAACATGCAGGCTCGACATTGGC----AAATTTTCTGGTTATCTTCAG------CTaTCTGGA	0.99
netA P2	AAGACTAATTACCATTTTCTCTCCT----TTTAGTCATTCTTATATTCTA----ACGTaGTCTTT	0.98
netBL	---TTACCGTGACATCGTGTAATGC----ACCTGTCGGCGTGATAATGCA---TATAatttTAAC	0.99
netF	--------TTTTCGGTTGACGCCCTT----CGGCTTTTCCTTCATCTTTac--------caTCTG-	0.99
nusA	--------CAGTATTTGCATTTTTT----ACCCAAAACGAGTAGAATTTG----CCACgTTTCAG	0.99
ppc	CGATTTCGCAGCATTTGACGTCACC----GCTTTTACGTGGCTTTATAAA-------AgaCGACG	0.99
pyrB1 P2	--TTGCATCAAATGCTTGCGCCGCT----TCTGACGATGAGTATAATGCC-------GgacAATT	0.99
pyrD	TTGCCGCAGGTCAATTCCCTTTTGG----TCCGAACTCGCACATAATACg--------ccccCG-	0.99
pyrE P1	ATGCCTTGTAAGGATAGGAATAACC----GCCGGAAGTCCGTATAATGCG------CAgCCACAT	0.99
rnh	GTAAGCGGTCATTTATGTCAGACTT----GTCGTTTTACAGTTCGATTca--------aTTACA-	0.96
rnp(RNaseP)	ATGCGCAACGCGGGGTGACAAGGGC----GCGCAAACCCTCTATACTGCG----CGCCgAAGCTG	0.99
rpmH1p	GATCCAGGACGATCCTTGCGCTTTA----CCCATCAGCCCGTATAATCCT------ccacccGGC	0.99
rpmH2p	ATAAGGAAAGAGAATTGACTCCGGA----GTGTACAATTATTACAATCCG--------gcctcTT	0.99
rpmH3p	AAATTTAATGACCATAGACAAAAAT----TGGCTTAATCGATCTAATAAA--------gatcCCA	0.99
rpoD Pa	--CGCCCTGTTCCGCAGCTAAAACG----CACGACCATGCGTATACTTT------AggGTT--	0.98
rpoD Phs	--ATGCTGCCACCCTTGAAAAACTG----TCGATGTGGGACGATATAGCA-----GAtaaG----	0.99
rrn4.5S	-GGCACGCGATGGGTTGCAATTAGC----CGGGGCAGCAGTGATAATGCG----CCTGCgCGTTGG	0.99
sdh P1	-ATATGTAGGTTAATTGTAATGATT----TTGTGAACAGCCTATACTGCC---GCCAGtCTCCG-	0.99
spot42r	ATTACAAAAAGTGCTTTCTGAACTG----AACAAAAAAGAGTAAAGTTAG----TCGGgTAGGGT	0.99
trpS	CTACGGCGAGGCTATCGATCTCAGC------CAGCCTGATGTAATTTATC---AGTCTaTAAATG	0.99
tyrT	TCTCAACGTAACACTTTACAGCGGC-----GCGTCATTTGATATGATGCG----CCCCgCTTCCC	0.99
uvrC	GCCCATTTGCCAGTTTGTCTGAACG----TGAATTGCAGATTATGCTGAT-------GatcaCCA	0.99

26 phage promoters used for testing

Name	sequences	Score
HuPc-1	---------AAATTTTGAAAAGTAACTTTATAGAAAAGAATAATACTGAA-AAGTCAATttGGT-	0.91
HuPc-2	------GGAACACATTTAAAAACCC----TCCTAAGTTTTGTAATCTATA-----AAGttAGCAA	0.98
HuPe	TACCAAAAAGCACCTTTACATTAAG----CTTTTCAGTAATTATCTTTTT-----AGTaAGCTAG	0.99
T7A1	TATCAAAAAGAGTATTGACTTAAAG----TCTAACCTATAGGATACTTAC----AGCCaTCGAGA	0.99
T7A2	ACGAAAAACAGGTATTGACAACATG---AAGTAACATCGCAGTAAGATACA----AATCgCTAGGT	0.99
T7A3	GTGAAACAAAACGGTTGACAACATG----AAGTAAACACGGTACGATGTA----CCACaTGAAAC	0.99
T7C	CATTGATAAGCAACTTGACGCAATG----TTAATGGGCTGATAGTCTTAT-----CTTaCAGGTC	0.99
T7D	CTTTAAGATAGGCGTTGACTTGATG----GGTCTTTAGGTGTAGGCTTTA-----GGTgTTGGCT	0.99
lambdaPR	TAACACCGTGCGTGTTGACTATTTT----ACCTCTGGCGGTGATAATGGT-----TGCaTGTACT	0.99
lambdaPRM	AACACGCACGGGTGTAGATATTTAT----CCCTTGCGGTGATAGATTTAA-----CGTaTGAGCA	0.99
lambdaPL	TATCTCTGGCGGTGTTGACATAAAT----ACCACTGGCGGTGATACTGAG-----CACaTCAGCA	0.99
lambdaPO	TACCTCTGCCGAAGTTGAGTATTTT----TGCTGTATTTGTCATAATGAC----TCCTgTTGATA	0.99
lambdaPR	TTAACGGCATGATATTGACTTATTG----AATAAAATTGGGTAAATTTGA----CTCAaCGATGG	0.99
lambdaPRE	TAGAGCCTCGTTGCGTTTGTTTGCA----CGAACCATATGTAAGTATTTC-----CTTaGATAAC	0.80
lambdaPI	CGGTTTTTTCTTGCGTGTAATTGCG----GAGACTTTGCGATGTACTTGA-----CACtTCAGGA	0.98

Table 1 (cont.)

```
434PR       AAGAAAAACTGTATTTGACAAACAA----GATACATTGTATGAAAATACA---AGAAAgTTTGTT 0.99
434PRM      ACAATGTATCTTGTTTGTCAAATAC----AGTTTTTCTTGTGAAGATTGG----GGGTaAATAAC 0.98
P22PR       CATCTTAAATAAACTTGACTAAAGA-----TTCCTTTAGTAGATAATTTA-----AGTgTTCTTT 0.99
P22PRM      AAATTATCTACTAAAGGAATCTTTA----GTCAAGTTTATTTAAGATGAC------TTaACTATG 0.98
P22ant      TCCAAGTTAGTGTATTGACATGATA----GAAGCACTCTACTATATTCTC-----AATaggTCCA 0.99
P22mnt      CCACCGTGGACCTATTGAGAATATA----GTAGAGTGCTTCTATCATGTC-----AATaCACTAA 0.99
phiXA       AATAACCGTCAGGATTGACACCCTC----CCAATTGTATGTTTTCATGCC-----TCCaAATCTT 0.99
phiXD       TAGAGATTCTCTTGTTGACATTTTA---AAAGAGCGTGGATTACTATCTG---AGTCCgATGCT- 0.99
phiXB       GCCAGTTAAATAGCTTGCAAAATAC---GTGGCCTTATGGTTACAGTATG-----CCCaTCGCAG 0.99
fdVIII      GATACAAATCTCCGTTGTACTTTGT-----TTCGCGCTTGGTATAATCGC------TGgGGGTCA 0.99
fdX         TCCTCTTAATCTTTTTGATGCAATT-CGCTTTGCTTCTGACTATAATAGA----CAGGgTAAAGA 0.99

40 mutant promoters used for testing

Name                                    sequences                          Score

araI(c)     --AGCGGATCCTACCTGGCGCTTTT-----TATCGCAACTCTCTACTGTT-TCTCCATaCCCGT- 0.99
araI(c)X(c) --AGCGGATCCTACCTGGCGCTTTT---TATCGCAACTCTCTACTATTTC---TCCATaCCCGTT 0.99
argCBH-P1/6 TTTGTTTTTCATTGTTGACACACCT------CTGGTCATAATATTATCAA---TATTCaTGCAGT 0.99
argCBH-P1/LL TTTGTTTTTCATTGTTGACACACCT------CTGGTCATGATATTATCAA---TATTCaTGCAGT 0.99
argE/LL13   CCGCATCATTGCTTTGCGCTGAAAC----AGTCAAAGCGGTTATATTCAT--------aTGCGGA 0.99
bioP98      TTGTTAATTCGGTGTAGACTTGTAA---ACCTAAAATCTTTTAAATTTGG-----TTTaCAAGTC 0.99
gal-P2/mut-1 -TAATTTATTCCATGTCACACTTTC------GCATCTTTGTTATACTATG-----GTTaTTTCAT 0.99
gal-P2/mut-2 -TAATTTATTCCATGTCACACTTTC------GCATTTTTGTTATGCTATG-----GTTaTTTCAT 0.99
IS2I-II     ---------ATGTCTGGAAATATAG----GGGCAAATCCACTAGTATTAA----GACtaTCACTT 0.99
lacP115     -TTTACACTTTATGCTTCCGGCTCG----TATGTTGTGTGGTATTGTGAG-------cggataac 0.99
lambdaci17  GGTGTATGCATTTATTTGCATACAT----TCAATCAATTGTTATAATTGT-----TATcTAAGGA 0.99
lambdacin   TAGATAACAATTGATTGAATGTATG----CAAATAAATGCATACACTATA-----GGTgTGGTTT 0.98
lambdaL57   -TGATAAGCAATGCTTTTTTATAAT----GCCAACTTAGTATAAAATAGC----CAACcTGTTCG 0.99
lpp/P1      --ATCAAAAAAATATTCTGAACATA---AAAAACTTTGTGTTATACTTGT-----AACgCTACAT 0.99
lpp/P2      --ATCAAAAAAATATTCTGAACATA---AAAAACTTTGTGTTATAATTGT-----AACgCTACAT 0.99
lpp/R1      --ATCAAAAAAATATTCACAACATA----AAAAACTTTGTGTAATACTTG----TAACgCTACAT 0.99
mac11       -CCCCCGCAGGGATGAGGAAGGTGG---TCGACCGGGCTCGTATGTTGTG-----TGGaATTGTG 0.99
mac12       -CCCCCGCAGGGATGAGGAAGGTCG---TCGGACGGGCTCGTATGTTGTG-----TGGaATTGTG 0.99
mac21       -CCCCCGCAGGGATGAGGAAGGTCG---ACCTTCGGGCTCGTATGTTGTG-TGGAATTgTGAGC- 0.94
mac3        -CCCCCGCAGGGATGAGGAAGGTCG----GTCGACCGCTCGTATGTTGTG-TGGAATTgTGAGC- 0.98
malPQA516p1 ---ATCCCCGCAGGATGAGGAGCCT------GGCAAACTAGCGATGATACG-TTGTGTTgAA---- 0.98
malPQA516P2 ATCCCCGCAGGAGGATGAGGAGCCT---GGCAAACTAGCGATAACGTTGT-----GTTgAA---- 0.95
malPQA517A  CCCCGCAGGATGAGGTCGAGCCTGG-----CAAACTAGCGATAACGTTAT-----GTTgAA---- 0.71
malPQ/Pp12  -ATCCCCGCAGGATGAGAAAGGTCA----ACATCGAGCCTGGAAAACTAG----CGATaACGTTG 0.98
malPQ/Pp13  -ATCCCCGCAGGATTAGAAAGGTCA---ACATCGAGCCTGGCAAACTAGC-----GATaACGTTG 0.99
malPQ/Pp14  -ATCCCCGCAGGATGAGAAAGGTCA----ACATCGAGCCTGGAAACTAGC-----GATaACGTTG 0.99
malPQ/Pp15  -ATCCCCGCAGGATGAGAAAGGTCA---ACATCGAGCCTGGCAAACTAGC-----GATaACGTTG 0.98
malPQ/Pp16  ATCCCCGCAGGATAAGGAAGGTCAA----CATCGAGCCTGGCAAACTAGC-----GATaACGTTG 0.96
malPQ/Pp18  ATCCCCGCAGGATGGGGAAGGTCAA----CATCGAGCCTGGCAAACTAGC-----GATaACGTTG 0.97
MR1rnaC/m   TCACAATTCTCAAGTTGCTGATTTC----AAAAAACTGTAGTATCCTCTG-------CgaaacGA 0.99
cmpF/pK1217 ------------GGTAGCGAAACGT----TAGTTTGCAAGCTTTAATGCG------GtaGTTTAT 0.99
pBRH313tet  ---AATTCTCATGTTTGACAGCTTA-----TCATCGATAAGCTAGCTTTAA-----TGCgGTAGTT 0.99
pBRH4-25    -----------TCGTTTTCAAGAAT-----TCATTAATGCGGTAGTTTATC------AcagTTA-- 0.99
pBRtet-10   AAGAATTCTCATGTTTGACAGCTTA----TCATCGATGCGGTAGTTTATC------AcagTTA-- 0.99
pBRtet-15   AAGAATTCTCATGTTTGACAGCTTA----TCATCGGTAGTTTATTACAGT------TAaatTGC- 0.99
pBRtet-22   AAGAATTCTCATGTTTGACAGCTTA---TCATCGATCACAGTTAAATTGC------TAacgCAG- 0.99
pBRtet/TA22 -----TTCTCATGTTTGACAGCTTA----TCATCGATAAGCTAAATTTTA----TATAaaATTTT 0.99
pBRtet/TA33 -----TTCTCATGTTTGACAGCTTA----TCATCGATAAGCTAAATTTAT-----ATAaaATTTT 0.99
pEG3503     -----------GGCTGGACTTCGAA---TTCATTAATGCGGTAGTTTATC------AcagTTA-- 0.99
TAC16       -----AATGAGCTGTTGACAATTAA-----TCATCGGCTCGTATAATGTG----TGGAaTTGTG- 0.99
```

a blank (−) instead of *X* and tested on 5000 random sequences (The random '−'*s* introduced were in the same positions where '*X*'*s* were introduced in the previous test). Though the prediction went down by 8.3% (93.3% to 85%), (Fig 2), the biasing was reduced considerably. This fall in prediction could be also because the random blanks introduced were in crucial positions in deciding between a promoter and a non- promoter. Hence, using blank (−) was a better choice.

In addition, the network was trained with non-promoters of two kinds, some entire 65 length sequences and the other with blanks introduced in it. While introducing blanks in the non-promoters, the similar distribution along the sequence like promoters (see text: alignment of promoters) were maintained.

DATA USED FOR THE STUDY

The learning set consisted of 106 promoters (3 (with 15 nucleotide (nt) spacer (sp)), 29 (with 16 nt. sp.), 41 (with 17 nt. sp.), 27 (with 18 nt. sp.), 4 (with 19 nt. sp.), 1 (with 20 nt. sp.) and 1 (with 21 nt. sp)). The network was tested on another set of 126 promoters. The promoters used for training and testing the network are shown in Table 1.

Table 2. Boxes identified differently by module I

Promoters	-35 Box		-10 Box		Spacer length	
	Ref 5	This Network	Ref 5	This Network	Ref 5	This Network
dnaA2p	AAGATC	CAGAAG	TATGAT	TATGAT	17	20
malPQ	AGGAAG	GAGGAA	CAAACT	CAAACT	17	18
		ATGAGG				20
P22PRM	AGGAAT	CTAAAG	TAAGAT	TAAGAT	17	21
malPQA5171A	GTCGAG	TCGACG	TAACGT	TAACGT	16	15
		ATGAGG				21

The promoters where the boxes were identified differently by the network with reference to Harley and Reynolds[5] is shown. Few of the boxes identified by the network are more conserved.

Table 3. Output values of testing single point mutations of P22ant promoter

	-35	Spacer	-10		Output
RU1204	TTGACA	TGATAGAAGCACTCTAC	TAGATT	promoter	0.91
RU1101	TTGACA	TGATAGAAGCACTCTAC	TACATT	promoter	0.88
RU1041	TTGACA	TGATAGAAGCACTCTAC	TATGTT	promoter	0.68
R-9AC	TTGACA	TGATAGAAGCACTCTAC	TATCTT	promoter	0.50
R-30AC	TTGACC	TGATAGAAGCACTCTAC	TATATT	promoter	0.99
R-30AT	TTGACT	TGATAGAAGCACTCTAC	TATATT	promoter	0.99
RU1147	TTGACA	TGATAGAAGCACTCTAC	TATACA	promoter	0.98
RU1099	ATGACA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.02
RU630	TTTACA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.03
RU612	TTGCCA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.10
RU483	TTGACA	TGATAGAAGCACTCTAC	CATATT	non-promoter	0.19
RU392	TTGACA	TGATAGAAGCACTCTAC	TTTATT	non-promoter	0.07
RE56	TTGACA	TGATAGAAGCACTCTAC	TATATC	non-promoter	0.10
R-35AC	CTGACA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.02
R-34TA	TAGACA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.02
R-33GC	TTCACA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.03
R-32AT	TTGTCA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.05
R-31CA	TTGAAA	TGATAGAAGCACTCTAC	TATATT	non-promoter	0.04
R-12TG	TTGACA	TGATAGAAGCACTCTAC	GATATT	non-promoter	0.16
R-11AC	TTGACA	TGATAGAAGCACTCTAC	TCTATT	non-promoter	0.09

Promoters have output values greater than 0.8 and non-promoters have output values less than 0.8. The first 7 sequences are experimentally found to be promoters while the others are reported to be non-promoters [12,13].

RESULTS AND DISCUSSIONS

The Neural networks in Module I, identified the two conserved boxes in all of the 126 tested promoters. But there were a few set of possible pairs of boxes identified in all promoters tested (8 sets on an average). This was because the network had learnt boxes from a wide variety of promoters which had the −10 and −35 regions conserved in only two positions to those conserved in all the six positions. Hence the network predicted boxes which were not highly conserved also. There were only 8 cases where the network identified boxes which were different from the ones suggested (5), with a cut-off value of 0.8. They were in the promoters; cat, dnaA 2p, groE, malPQ, lambdaPRE, P22PRM, malPQA517A and malPQ/Pp16. In both cat and lambdaPRE, the second set of boxes predicted in (5) were identified. In dnaA2p, malPQ, P22PRM, malPQA517A and malPQ/Pp16, with the cut-off at 0.8, different boxes were identified. The different boxes identified are shown in Table 2. If the cut-off is changed to 0.5, the exact boxes specified (5) could also be identified.

The module III of the network was tested on 60 bacterial, 26 bacteriophage and 40 mutant promoters mentioned in Table 1. Sequences that had an output of more than 0.8 were considered as promoters and an output below 0.8 as non-promoters. All the 126 promoters tested, except two (nanA and malPQA517A),

were identified as promoters. Figures 3a and 3b show the percentage of promoters and non-promoters predicted respectively at various output values.

Identifying single point mutations in P22ant promoter

Depending on the position, a point mutation may or may not convert a promoter into a non-promoter. The predictive ability of the network described here can be tested if it can identify the promoter and non-promoter sequences of the 39 known point mutations of the P22ant promoter. The possible mutations of the P22ant promoter and their effect on promoter activity have been studied (12,13). There were 39 mutations of the promoter P22ant done mainly in the −35 and −10 regions. Of the 39 mutations, 3 of them were two-point mutations ,one was a deletion ,while the rest were single point mutations. 24 of the mutations resulted in sequences which were non-promoters and 15 of them remained as promoters. Along with the 106 promoters and 212 non-promoters, a few of the promoters and non-promoters obtained by mutating the P22ant promoter were also taught to the network.

The ratio of promoters to non-promoters (obtained by mutating the P22ant promoter) taught to the network was varied. This did not alter the prediction very much. When the set of non-promoters taught to the network was varied by taking a few non-promoters at a time for training and was tested on the remaining, the prediction capabilities changed. Since varying the non-promoter mutated set for training the network affected the prediction, a set of eight 'worst' mutated non-promoters were chosen for training. 'Worst' mutated non-promoters are defined as those non-promoters which were obtained by a mutation which resulted in a base change from a high probability of occurrence to a base with a low probability of occurrence . In this case since the network could learn the information that made the sequence a non-promoter, it was able to predict better.

The network learnt 11 mutated non-promoters (RU369, RU454, RE167, R204, RU267, RU523, RU541, RU428,R-34TG,R-31CG and R-7TG) and 8 mutated promoters (RU1150, RU287, RU1002, RU1156, RU1012, R1173, RU1197 and R-30AG) along with P22ant. The network was tested on the remaining 13 mutated non-promoters and 7 mutated promoters. The output values obtained are shown in Table 3. All the promoters (except two, RU1041 and R-9AC)and non-promoters were predicted correctly by the network. This shows that the network can identify single point mutations also if it is taught typical promoters and non-promoters which carry this information.

Building up of a general network to predict the promoters in plasmid sequences

The above described network ,specially in module III , had a very limited training set (106 promoters and 212 non-promoters) compared to the number of free parameters (from the 1827 weights and 8 bias , ie total 1835) . Hence it might not have had sufficient information for predicting promoters in the general sequences . It was observed that, in the plasmid pBR322 too many false positives were predicted ,which was due to this limitation. To correct this feature within the available data (i.e., only 419 *E.coli* promoter sequencs available (ref.5,14)), 369 promoters and 1296 non-promoters (identified as described earlier) were used to train the general network and tested on the rest 50 promoters and 150 non-promoters to judge the performance of the network. We have also used 153 unique promoter boxes at −10 and 207 unique promoter boxes at −35 in module I. It was found that prediction of non-promoters increased to 98.7% and of promoters to 98% .

Table 4. Sequences of known promoters in pBR322, identified by the neural network

	-35	Spacer	-10	SL	Homology Score	Neural net Output
P1	CTGACT	GCGTTAGCAATTTAACTGTGA	TAAACT	21	58.6	0.99
P2	TTGACA	GCTTATCATCGATAAGC	TTTAAT	17	63.9	0.99
P3	TTCAAA	TATGTATCCGCTCATGA	GACAAT	17	52.6	0.99
P4	TTGAAG	TGGTGGCCTAACTACGGC	TATACT	18	64.5	0.99
P5	CTGAGA	GTGCACCATATGCGGTG	TGAAAT	17	46.74	0.99

The table gives the possible−35, spacer and −10 regions of the P1, P2, P3, P4 and P5 promoters in pBR322 sequence with the spacer length (SL), homology score (obtained by TARGSEARCH) and the output value of the neural network. P2, P4, P5 are on the reverse strand of pBR322.

Predicting the promoters in pBR322 sequence

The above described network was used to identify all the promoter sequences present in plasmid pBR322. The identification was done on both the pBR322 sequence and the reverse sequence of the plasmid. The network was used to identify the known promoter sites of P1, P2 ,P3,P4 and P5 on pBR322 (15, 16).

The pBR322 sequence (length 4363 bases) and the reverse pBR322 sequence were divided into smaller 65 base sequences using a window method with an overlap of 60 bases. The windows were generated starting from the 5′ end of the sequence. Each of these smaller sequences was passed through module I of the network. All purines present in the first 10 bases from the 3′ end towards the −10 box were considered as possible initiation codon points and the promoter boxes in all the sequences were identified. Module II aligned the sequences according to the boxes recognized. These aligned sequences were passed through module III and the sequences with output more than 0.8 were considered promoters. Since an overlap of 60 was used in creating windows of length 65, the duplicates of promoters recognised in more than one window were eliminated. The network identified 371 unique promoter sequences in the pBR322 sequence and 393 promoters in the reverse strand of pBR322. The reason for recognising many promoters was that the network had learnt a variety of promoters with highly conserved to weakly conserved −35 and −10 boxes and hence it identified promoters of all strengths. Since the initiation point for each window was not known, all purines in the first 10 bases from the 3′ end towards the −10 box were considered as starting points which also resulted in finding more promoters. The TARGSEARCH program (11) also found as many as 1396 promoters in the pBR322 plasmid and its reverse strand with a match of 3 out of 6 in both the conserved regions. The homology scores for all the predicted sequences were found using the TARGSEARCH program (11) also . This network has identified the possible boxes in all the five cases, P1, P2, P3, P4 and P5 known promoters in PBR322, within the specified region. The predicted sequences of −35 box, Spacer and −10 box of all the five promoters with their homology scores found by TARGSEARCH and neural network output are shown in Table 4. The score of the promoter P5 in the pBR322 sequence was 46.7 which was the lowest among the five identified promoters (11) and the highest was for P4 (64.5) which is known to be a strong promoter (15).

Table 5a. Promoters with start codon and ribosomal binding site in pBR322 sequence

-35	Spacer	-10	Sep	RBS	Gap	St	Scr	Pr
CTGCGT	TAGCAATTTAACTGTGA	TAAACT	242	AGGA	5	ATG	47.34	Y
CTGACT	GCGTTAGCAATTTAACTGTGA	TAAACT	242	AGGA	5	ATG	58.58	Y
TTCACA	CCGCATATGGTGCACTCTCAG	TACAAT	157	GGAG	6	GTG	47.34	Y
CTGTCA	GACCAAGTTTACTCATA	TATACT	34	AGGA	5	GTG	54.43	N
TGGAGC	CGGTGAGCGTGGGTCTCGCGG	TATCAT	193	AGGA	5	GTG	47.34	N
TTAATA	GACTGGATGGAGGCGGA	TAAAGT	283	AGGA	5	GTG	45.56	N
TTAAAG	TTCTGCTATGTGGCGCGG	TATTAT	256	GGAG	4	ATG	46.15	N
TTGAGT	ACTCACCAGTCACAGAA	AAGCAT	164	GGAG	4	ATG	46.75	N
TTCAAA	TATGTATCCGATCATGA	GACAAT	32	AGGA	5	ATG	52.66	Y
TTGTTT	ATTTTTCTAAATACAT	TCAAAT	54	AGGA	5	ATG	45.46	Y
ATGTCA	TGATAATAATGGTTTAT	TAGACG	123	AGGA	5	ATG	50.29	Y
TTTTTA	TAGGTTAATGTCATGAT	AATAAT	136	AGGA	5	ATG	49.11	Y
GTGATA	CGCCTATTTTTATAGGT	TAATGT	148	AGGA	5	ATG	51.47	Y

Table 5b. Promoters with start codon and ribosomal binding site in reverse strand of pBR322 sequence

-35	Spacer	-10	Sep	RBS	Gap	St	Scr	Pr
TTGACA	GCTTATCATCGATAAGC	TTTAAT	25	-	-	ATG	63.90	Y
TGGTCT	TCGGTTTCCGTGTTTCG	TAAAGT	279	GGAG	7	GTG	55.62	Y
CTGAGA	GTGCACCATATGCGGTG	TGAAAT	153	AGGA	6	GTG	46.74	N
GTGAAA	TACCGCACAGATGCGTAAGGA	GAAAAT	127	AGGA	6	GTG	48.52	N
TTGAAG	TGGTGGCCTAACTACGGC	TATACT	211	AAGG	10	ATG	64.50	N
TTCACC	TAGATCCTTTTAAATT	AAAAAT	241	AAGG	13	GTG	50.89	N
TTCACC	TAGATCCTTTTAAAT	TAAAAA	242	AAGG	13	GTG	49.70	N
ATGAAG	TTTTAAATCAATCTAAAG	TATATA	213	AAGG	13	GTG	45.56	N
TTAAAT	CAATCTAAAGTATATATGAG	TAAACT	203	AAGG	13	GTG	53.25	N
TTTAAA	TCAATCTAAAGTATATATGAG	TAAACT	203	AAGG	13	GTG	53.57	N
CTGACT	CCCCGTCGTGTAGATAAC	TACGAT	96	AAGG	13	GTG	52.07	N
CTGTCA	TGCCATCCGTAAGATGC	TTTTCT	260	AGGA	7	ATG	46.15	N
TTCTCT	TACTGTCATGCCATCCG	TAAGAT	268	AGGA	7	ATG	53.84	N

Table 5a and 5b show the −35 region, spacer and −10 regions of promoters with the possible start codon (St) and Ribosomal Binding Site (RBS). The distance between −10 box and RBS is shown in column labelled Sep. The number of bases between the start codon and RBS are shown in the column labelled gap. Homology scores by TARGSEARCH are provided in the column labelled Scr. The promoters with more than 25 amino acids separating the start and stop codon are represented in the column labelled Pr with a Y and the rest are represented using N. In 5a, the second promoter is P1 and 9th is P3 , whereas, in 5b, the first one is P2, 5th. is P4 and 3rd. is P5 . In Table 5b, from 6th. to 11th. promoter sequences transcribe for the same RNA.

Table 6a. Promoters in the forward strand of plasmid pWM528

-35	Spacer	-10	Sep	RBS	Gap	St	Scr	Pr
TTTACA	CTTTATGCTGCCGGCTCG	TATGTT	33	AGGA	7	ATG	49.7*	Y
TTTGTT	TATTTTTCTAAATACAT	TCAAAT	54	AGGA	5	ATG	46.1	N
TTCAAA	TATGTATCCGCTCATGA	GACAAT	32	AGGA	5	ATG	52.7*	Y
TTGAAA	AAGGAAGAGTATGAGTATTC	AACATT	87	AGGA	16	ATG	50.9	Y
TTGAAT	AGCGGTAAAATCCTT	GAGAGT	54	GGAG	12	GTG	45.6	N
TGGACT	TGAATAGCGGTAAAATCCTT	GAGAGT	54	GGAG	12	GTG	46.7	N
TTGAAT	AGCGGTAAAATCCTTGA	GAGTTT	52	GGAG	12	GTG	51.5	N
GTGCCG	CCATAACGATGAGTGAT	AACACT	66	GGAG	3	ATG	46.1	N
TTAATA	GACTGGCTTGAAGCGGA	TAAAGT	63	AGGA	4	GTG	45.6	N
CTGAAA	TAGGGGCTTCACTGATT	AAGCAT	-	-	-	-	47.3	N
TTCACA	CAGGAAACAGCTATGAC	TATGAT	-	-	-	-	45.0	N
CTGATT	AAGCATTGGTAAACCGA	TACAAT	-	-	-	-	46.1	N
TTCACT	GATTAAGCATTGGTAAACCGA	TACAAT	-	-	-	-	55.6	N
TTGAGA	TCCTTTTTTTCTGCGCG	TAATCT	-	-	-	-	60.3*	N
TTGAGA	TCCTTTTTTTCTGCGCGTAA	TCTGCT	-	-	-	-	45.6	N

Table 6b. Promoters in the reverse strand of plamid pWM528

-35	Spacer	-10	Sep	RBS	Gap	St	Scr	Pr
TTGAAG	TGGTGGCCTAACTACGGC	TACACT	-	-	-	-	64.5*	N
CTGACT	ACCCGTCGTGTAGATAAC	TACGAT	285	AAGG	8	ATG	52.1	N
TTCTCT	TACCGTCATTCCGTCCG	TAAGGT	268	AGGA	7	ATG	50.9	N
TTGAAT	ACTCATACTCTTCCTTTTTC	AATATT	154	AAGG	7	GTG	52.7	Y
TTGAAT	ACTCATACTCTTCCTTTTT	CAATAT	155	AAGG	7	GTG	49.7	Y
TTGAAT	ACTCATACTCTTCCTTTT	TTCAAT	157	AAGG	7	GTG	56.2	Y
TACTCA	TACTCTTCCTTTTTCAA	TATTAT	152	AAGG	7	GTG	49.7	Y
TTGAAG	CATTTATCAGGGTTATTG	TCTCAT	123	AAGG	7	GTG	53.2	Y
TTGTCT	CATGAGCGGATACATAT	TTGAAT	103	AAGG	7	GTG	52.1	Y

Table 6a and 6b show the −35 region, spacer and −10 regions of promoters with the possible start codon (St) and Ribosomal Binding Site (RBS). The distance between −10 box and RBS is shown in column labelled Sep. The number of bases between the start codon and RBS are shown in the column labelled gap. Homology scores by TARGSEARCH are provided in the column labelled Scr. The promoters with more than 25 amino acids separating the start and stop codon are represented in the column labelled Pr with a Y and the rest are represented using N. If both RBS and Start codon are not been found, the corresponding columns are filled with '−'. In Table 6b, the last six promoters transcribe from the same start codon.

It has been stated by Mulligan *et al*, (11) that, the promoters identified by TARGSEARCH with a score below 45 would be functionally insignificant. The network identified 26 promoters with homology scores (by TARGSEARCH) more than 45 in the pBR322 plasmid and 22 promoters with scores greater than 45 in the reverse strand of pBR322. Mulligan *et al* have also reported 26 promoters with scores more than 47.3.

We have also added another module based on a string search procedure which identifies the promoters that preceed a translable gene region containing a ribosomal binding site (RBS) and a start codon. For all the promoters with homology scores more than 45, a search for start codon (ATG or GTG) within 300 bases from the start +1 point towards the direction of gene translation and a search for the RBS (AGGA or AGGT or AAGG or GGAG or GAGG) with 1 to 16 bases separating it from the start codon sequence towards the +1 point was done. A search for the stop codon (TAA or TAG or TGA) within 25 amino acids away from the start codon was also done. Out of the 26 sequences identified in the pBR322 strand (homology score greater than 45), 13 of them did not contain either the RBS or the Start codon. The remaining 13 promoters contained the promoters , P1 and P3. The −35 box, Spacer and −10 box of the possible promoters with the number of bases between −10 box and RBS and the number of bases between and RBS and Start codon are shown in Table 5a. The promoters with less than 25 amino acids separating the start and stop codon are also represented in the table. In the reverse strand of the pBR322 plasmid, 9 of the total 22 promoters with homology scores above 45 were eliminated similarly. Of the remaining 13 promoters, P2,P4 and P5 were also identified as shown in Table 5b.

The network was also tested on a synthetic plasmid pWM528 (17). The plasmid has 1993 bases. The network identified all the 3 known promoter sites on the forward strand and 1 known promoter site on the reverse strand of pWM528. The network identified 206 promoters in the forward strand and 209 promoters in the reverse strand of pWM528. The promoters with a homology score greater than 45 in the forward and reverse strand are shown in Table 6a and 6b respectively. The promoters identified in pWM528 by Mandecki *et al* are indicated by * .

CONCLUSIONS

1. A combination of two networks, one which has learnt the boxes and the other which has learnt the entire sequence can be used to predict promoters of all spacer lengths instead of using separate networks for the different spacer lengths.

2. Since the second network uses the information of the entire sequence, position by position and also by cross correlations of every position with every other position, the possible dependencies between the bases in various positions in the sequence which may influence the promoter prediction are taken into account.

3. In cases where the starting point and spacer length of the sequences are known, the module III of the network can be used

for identification. But in cases where the sequence is totally unknown, module I of this network finds boxes with the required spacer lengths taking any purine in the initial 10 bases from the 3′ end towards the −10 box as the starting point. This is aligned and tested on the network on module III for finding a promoter. Therefore one could use this method to predict promoters on any DNA sequence which may have a 'promoter-like' structure.

4. This network is able to identify single-point mutations.

REFERENCES

1. A.V.Lukashin, V.V. Anshelevich, B.R.Amirikyan, A.I. Gragerov and M.D.Frank-Kamenetskii, (1989), J. Biomol. Struc. Dynam., 6, 1123−1133.
2. B.Demeler and G.Zhou, (1991), Nucl. Acids Res. 19, 1593 - 1599.
3. M.C.O'Neill, (1991), Nucl. Acids Res., 19, 313 - 318.
4. M.C.O'Neill, (1992), Nucl. Acids. Res., 20, 3471 - 3477.
5. C.B. Harley and R.P.Reynolds, (1987), Nucl.Acids Res., 15, 2343−2361.
6. M.C.O'Neill, (1989), J.Biol. Chem., 264, 5522 - 5530.
7. L.R.Cardon and G.D.Stormo, (1992),J. Mol.Biol., 223, 159 - 170.
8. K.Hawley and William R.McClure, (1983), Nucl. Acids. Res., 11, 2237−2255.
9. U. Deuschle, W. Kammerer, R. Gentz and H. Bujard, (1986), EMBO J, 5, 2987−2994.
10. D.E.Rumelhart, G.E.Hinton and R.J.Williams, (1986), Nature,323, 533−536.
11. M.E.Mulligan, D.K. Hawley, R.Etriken and W.McClure, (1984), Nucl. Acids Res., 12, 789−800.
12. P. Youderian, S. Bouvier and M.M. Susskind, (1982), Cell, 30, 843−853.
13. H.Moyle, C.Waldburger and M.M.Susskind, (1991), J. Bact. 173, 1944 - 1950.
14. S.Lisser and H.Margalit , (1993), Nucl. Acids. Res., 21, 1507 - 1516.
15. D. Stuber and H. Bujard, (1981), Proc. Natl. Acad. Sci, USA, 78, 167−171.
16. J. Brosius, R. L. Cate and A. P. Perlmutter *et.al* (1982), J. Biol. Chem., 257, 9205−9210.
17. W. Mandecki, M.A. Hayden, M Ann Shallcross and E. Stotland, (1990), Gene, 94, 103−107.