Mimicking Transcription Process to Recognise Promoters in E.coli

T.Sobha Rani
Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India

Abstract— *Promoter prediction is a computationally interesting and complex problem. Various groups have tried promoter prediction with different sequential and structural features of promoters. The structural aspects of DNA in promoter recognition are gaining popularity of late. First step in transcription process is the binding of RNA polymerase with the promoter. Here in this work, a preliminary study of interactions between RNA polymerase and specifically the binding sites within the promoter is carried out. Interaction values between RNA polymerase and DNA are used to identify the -35 and -10 binding sites in the promoter. A set of windows around these regions are extracted. Bi-gram features of these windows are used to test the validity of using such interactions in promoter recognition. Two types of encoding, Electron-ion interaction potential (EIIP) and amino acid-base pair inetraction values are used to quantify the interaction between RNA polymerase and the promoter. Current results are comparable to earlier results obtained with n-grams. The experiments seems to point to a signal global in nature is much more efficient than local signal in promoter recognition. The results also confirm that the basic interactions between RNA polymerase and DNA (promoter) have the capability to identify the promoters in a whole genome.*

Keywords: Classification ; EIIP encoding ; amino acid-base pair integration; machine learning

1. Introduction

Promoter prediction is complex and several groups of researchers have attempted to solve this problem by extracting different features which can be used to characterize the promoters. Some of the features that have been used for this task are position weight matrices [1], [2], [3], n-mers [4], [5], [6] which are statistical in nature. There are methods that have used DNA structural features such as enthalpy [7], thermal stability [8], stress induced duplex destabilization [9], roll-angle [7], base stacking energy [10] etc. Ponomarenko et al. have listed a wide variety of structural properties [11]. A wide range of classifiers such as neural networks [13], [1], SVM [12], hidden Markov model [14] and graph based induction [15] are also used.

Even though there is a huge amount of work done, the promoter prediction problem is far from being solved. The accuracy of predictions is not very high. In case of eukaryotes a group of promoters called GC rich promoters are easier to predict than other promoters which are not GC rich. We want to investigate this problem from the point of view of the basic chemical interactions that arise between the RNA polymerase and the promoter irrespective of the nature of the promoters present in the genome. As a consequence, DNA-RNA polymerase interactions and bi-grams are used in the promoter identification in this work.

1.1 DNA-RNA Polymerase Interaction

In prokaryotes, the first step in transcription is the binding of RNA polymerase with the promoter. RNA polymerase is a large molecule consisting of five subunits α_1, α_2, β, β' and ω. In order to bind promoter-specific regions, the core enzyme requires another subunit, sigma (σ). The sigma factor greatly reduces the affinity of RNAP for nonspecific DNA while increasing specificity for certain promoter regions, depending on the sigma factor. This way, transcription is initiated at the right region. The complete holoenzyme therefore has 6 subunits: $\alpha_1\alpha_2\beta\beta'\omega\sigma$ (480 kDa). The structure of RNAP exhibits a groove with a length of 55 (5.5 nm) and a diameter of 25 (2.5 nm). This groove fits well the 20 (2 nm) double strand of DNA.

Promoter specific transcription on RNA polymerase is conferred by σ subunit. Based on sequence analysis these σ factors are divided into two broad classes σ-70 factors and σ-54 factors. Four highly conserved regions are identified by aligning σ70 family of proteins [16], [17], [18]. Of these regions 2 and 4 are highly conserved and basic in nature and regions 1 and 3 exhibit low conservation and are acidic in nature. The secondary structures of regions 1 and 2 are predicted to be β-sheets with helices and regions 3 and 4 are predicted to be helical [19].

A series of studies revealed that sub-region 2.4 (located at the C-terminal end of region 2) interacts directly with promoter -10 hexamer elements, whilst sub-region 4.2 (located at the C-terminal end of region 4) interacts directly with promoter -35 hexamer elements. A number of studies using a variety of primary and alternative σ factors from E.coli and B.subtilis have identified residues of region 2.4 (a sub region of region 2) interacting with -10 hexamer and these interactions are depicted in figure 1 [20], [21], [22], [23]. Genetical analysis studies explain the interactions between the residues of RNA polymerase and nucleotides of -35 region in DNA [24], [25]. Figure 2 illustrates these interactions between the residues of $\sigma 4.2$ region and the -35 region of the promoter. Eventhough a lot of other interactions

are involved, only the interactions between RNA polymerase and the binding sites is considered here as a starting point.

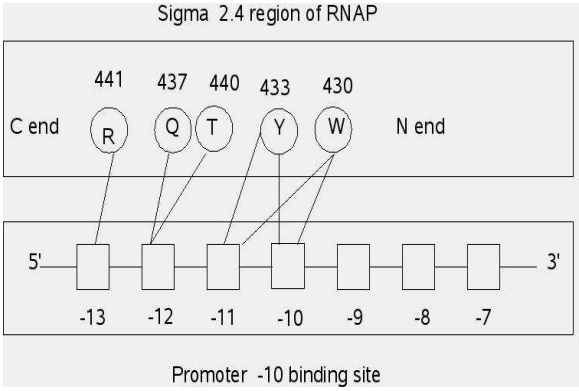

Figure 1: Pictorial depiction of the interactions between -10 binding site and amino acids of σ subunit.

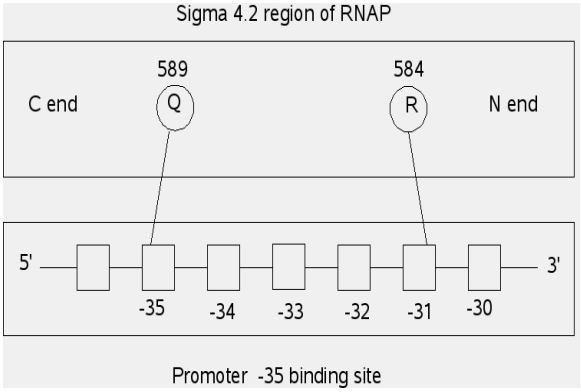

Figure 2: Pictorial depiction of the interactions between -35 binding site and amino acids of σ subunit.

A systematic study of n-grams in promoter prediction for $n = 2, 3, 4, 5$ [6] was carried out by us. We have obtained 68% promoter prediction accuracy for E.coli with $n = 3$. We got a very good prediction of promoters on forward-strand of E.coli taken from NCBI data base [6].

The main difference between the work that is being proposed in this paper and the work reported by Sobha et al. [6] is that in this paper emphasis is on identifying the binding sites through interaction between the DNA and RNA polymerase whereas in the later work it is just the occurrence of n-grams in the whole promoter without distinguishing the binding sites and non-binding sites.

2. Approach

A preliminary study of DNA-RNA polymerase interaction information in promoter recognition is performed by us [26]. We have attempted to compute the interaction through cross-correlation between promoter and RNA polymerase sigma subunit. We have not considered the three-dimensional aspect of RNA polymerase then. Hence, the results of classification were not good for promoters. Here, in this paper we have tried to identify a subset of amino acids in RNA polymerase sigma subunit that takes part in the interaction between promoter and RNA polymerase. These appear mostly as part of α helix in $\sigma 2$ and $\sigma 4$ regions of σ subunit.

Interaction between RNA polymerase and promoter is quantified in two ways. One is by computing the cross-correlation between the DNA and RNA polymerase signals converted into numerical sequences. Second one is by considering the the values obtained considering the interaction between amino acids and nucleotides. Cross-correlation between the residues of sigma subunits of RNA polymerase which interact with -10 and -35 hexamer regions are converted into numerical sequence using the EIIP values for the amino acid [27]. Similarly the nucleotides which take part in this interaction are also converted into numerical sequences using EIIP encoding [28]. Since we have no knowledge about the nucleotides which interact with the amino acids in a sigma subunit, we have followed the spacer scheme of Ma et al. [13]. They have considered a varying space of 15-21 bp (7 bp) between -35 and -10 regions and 3-11 bp (9 bp) between -10 region and TSS. In the same way, we have constructed our sigma subunit segment of length 80, consisting of zeroes except at the positions -35, -31 and -13, -12, -11, -10 positions with different spacings between them. This would result in a set of 63 combinations. Maximum correlation coefficient of the 63 combinations is chosen to fix the spacers between -35 and -10 regions and also between -10 region and TSS. Once the spacers are fixed, we can identify the binding regions in a promoter. Windows of certain length are extracted around these binding sites. Bi-gram features of these windows are extracted as features for a multi-layer feed forward neural network to train and identify the promoters in a genome.

3. Methodology

E.coli promoter data set is used for experimentation. We consider sequences of length 80 bp with 60 base pairs upstream of the Transcription Start Site (TSS) and the rest downstream [12]. Positive data set consists of 669 promoter sequences of length 80 bp [12]. Negative data sets of Gordon et al. who have chosen these in a biologically meaningful way by taking sequence fragments outside the promoter region. They also have built negative data sets with 709 sequence fragments from coding region and 709 sequence segments from intergenic portions.

3.1 Feature Extraction

Features are extracted in two stages. In the first stage, DNA-RNA polymerase interaction is used. In the second stage, windows around binding sites are identified and bi-gram features of these are extracted. These features are used

as input for a multi-layer feed forward network to learn about promoter. Here, the promoter recognition is posed as a binary classification problem.

3.1.1 Step 1: Identification of binding sites using DNA-RNA polymerase interaction

From literature the amino acids that interact with -10 and -35 binding regions are identified. The residues that interact with -35 binding site are taken from the work of Campbell et al. [25]. Similarly the residues that participate in interaction with -10 binding site are taken from the work of Malhotra et al. [23]. These are depicted in Figure 1.

In order to compute the interaction between DNA and RNA polymerase, we have chosen the cross-correlation as the means. Cross-correlation between the two can be computed by converting both DNA and RNA polymerase sequences into numerical sequences. In this method the amino acid residues and nucleotides are encoded into numerical format using EIIP values [27], [28]. EIIP encoding is chosen since it can be used to encode both amino acids and nucleotides. Table 1 lists the EIIP values of the relevant amino acids and nucleotides.

Table 1: EIIP values for amino acids [27] and nucleotides [28].

Amino acid	EIIP	Nucleotide	EIIP
Tyrosine(Y)	0.0516	A	0.1260
Tryptophan(W)	0.0548	T	0.1335
Glutamine(Q)	0.0761	G	0.0806
Threonine(T)	0.0941	C	0.1340
Arginine (R)	0.0959		

Another way of encoding using values provided by Mandel et al. [29]. Mandel et al. [29] have analyzed protein-dna complexes to extract all non-homologous pairs of amino acid-base pairs that are in close contact. A quantitative measure of the likelihood of the interaction between each pair of amino acid and base is computed. A score can be computed by summing up the individual measures of amino acid-base pairs assuming additivity in their contributions to binding. This score can be used a measure of the compatibility between the protein and its dna target. Table 2 lists these amino acid-base pair interaction values.

Table 2: Amino acid-base pair interaction values [29].

	G	A	T	C
Trp	-1.96	-3.93	-1.96	-3.93
Tyr	-2.87	-2.87	0.54	0.13
Gln	-0.09	1.16	0.31	-3.09
Thr	-3.46	-0.06	-0.06	-1.16
Arg	2.74	0.34	1.25	-3.93

In order to obtain the interaction between promoter and residues in the sigma subunit, $similar(j)$ defined in 1 is computed.

$$similar(j) = min(\Sigma abs(s_1 - s_2)); j = 1, 2, 3, \ldots, 63 \quad (1)$$

Here j=1 denotes the spacing between -35 and -10 regions as 15, and spacing between -10 and TSS as 3 bp. Similarly, j=2 denotes spacing between -35 and -10 regions as 16, and spacing between -10 and TSS as 3 bp and so on. Final j=63 denotes spacing between -35 and -10 regions as 21, and spacing between -10 and TSS as 11 bp. These are listed in Table 3. $similar(j)$ will be close to zero if s_1 and s_2 are close to each other. That is, if we suppose s_1 as promoter and s_2 as the set of residues that interact with the promoter, when they are compatible with each other, then $similar(j)$ values will also be zero.

Table 3: Spacing between -35 and -10 binding sites (SP35) and -10 and TSS (SP10) for different j values.

j	SP35 (bp)	SP10 (bp)
1	15	3
2	16	3
..	..	..
7	21	3
8	15	4
..	..	..
63	21	11

The values of the 63 combinations for various spacings between -35 and -10 regions and -10 and TSS can be treated as the compatibility between the sigma subunit and promoter binding regions. Of the 63 combinations obtained from the above calculations, the highest score is considered to arrive at the spacers between binding sites. Fixing up the spacers, binding sites can be identified.

3.1.2 Step2: Bi-gram feature extraction

Regions with high information content are selected, specifically 17 positions around the -35 binding site and 11 positions around the -10 binding site and 7 positions around the transcription start site are extracted. A bi-gram is a combination of contiguous two letters. DNA consist of four bases and therefore 16 bigrams (AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC) are formed [6]. For each window 16 bigrams are computed. In total 48 bi-gram features are obtained for all the three windows. Two types of experiments are performed. One in which the original 48 bigram features are given as input features to the multi-layer feed-forward (MLFF) perceptron. The output of the neural network is ≥ 0.5 if the given sequence is predicted as a promoter. Second one in which the bi-gram feature values for each of the windows are combined together into 16 bigram features. Simulations are done using SNNS package [30].

3.2 Training and Testing

Bi-gram features extracted from the windows around the binding sites and TSS are used as input features for the MLFF neural network. We have carried out 5-fold cross-validation procedure in which the total data set is divided into 5 parts. In each fold, 1 part will form the test set while the remaining four will be used for training. Precision (Pr), Sensitivity (Sn) and Specificity (Sp) are used as measures of classification performance. Specificity is the proportion of the negative test sequences that are correctly classified and sensitivity is the proportion of the positive test sequences that are correctly classified. Precision is the proportion of the correctly classified sequences of the entire test data set.

3.3 Extension to Whole Genome Promoter Prediction

The real test for any promoter recognition is it's ability to identify promoters in a whole genome. Towards this end, we have used $section1$ and $section3$ of E.coli. Total genome of E.coli is divided into 400 sections. Out of these sections two sections $section1$ and $section3$ are chosen to extend the promoter recognition algorithm. These are chosen for the purpose of comparison with the results obtained using n-gram features [6]. A sliding window of 80 bp is used to segment these sections into segments of size 80 bp. We consider a sliding window of length 80 extracting segments from the start of the DNA sequence considered, that is, 1–80, 2–81, 3–82 and so on. These are represented as the bi-gram feature vectors which are used by the neural network classifier. Each of the segments gets classified as promoter (P) or non-promoter (NP). If a segment $m-(m+79)$ is classified as a promoter, then the nucleotide m is annotated as P and if it is classified as non-promoter then m is annotated as NP. This process of annotation is continued for the entire sequence to get a sequence of P's and NP's. We propose that if a contiguous segment of length more than a certain threshold has all P's then we annotate that region as promoter region otherwise as non-promoter region. For the verification purpose we have considered the $section1$ and $section3$ of E.coli [31]. It also denotes the set of promoters present in these segments.

4. Discussion

Table 4 shows the average of 5-fold cross validation results for both 48 bigram features as well as 16 bigram features extracted using $similar(j)$ (refer to equation 1). Sensitivity and specificity using 48 bigram-features are close to what was obtained using bi-grams for the entire promoter [6] than the ones obtained with 16-bigram features. But $section1$ and $section3$ results using 48 features and 16 features extracted from the windows using $similar(j)$ values present a different scenario. False-positives, that is non-promoters identified as promoters are much less with 16 features compared to 48 features. This fact is evident from figures 3 and 4. In these figures X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window. Only output greater than 0.5 is considered as a promoter. It is not only an output greater than 0.5 that is essential we also need to have a stretch of continuous ones over a threshold value of 20-25 is required to annotate the stretch of base pairs as as a promoter. In this context, we could identify a clear stretch of positives with 16 features compared to 48 features. These results are comparable to what was obtained using n-grams [6]. We have also carried out one more experiment wherein the nucleotides in the windows are straightaway used as input features to the neural network after converting them into numerical values using EIIP codes. This experiment results are not as good as the results obtained with bi-grams.

Table 4: Classification results using $similar(j)$features. SetA: Bi-grams from each window used separately and SetB: Bi-grams from each window combined together.

Features	Number	Pr	Sp	Sn
SetA	48	79.15	86.92	62.76
SetB	16	76.47	86.53	55.14

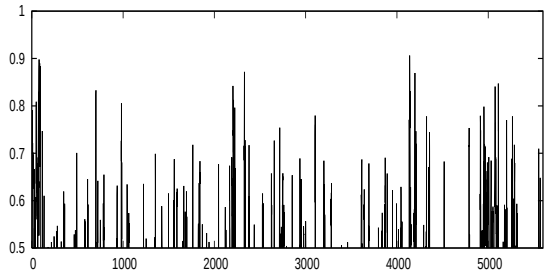

Figure 3: $section1$ of E.coli tested with 48 bi-gram features. X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window.

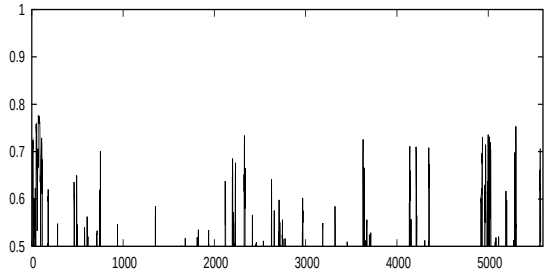

Figure 4: $section1$ of E.coli tested with 16 bi-gram features. X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window.

As in the case of n-grams extracted from the whole pro-

moter, we have obtained satisfactory results with the features extracted from the interaction between promoter and certain residues of sigma subunit. Through this interaction, we have extracted the binding sites and the windows around the binding sites and TSS. Whole genome promoter prediction results using 16 bi-gram features in fact assures that the binding sites that are extracted are of relevance since we obtain similar results as in the case of bi-grams extracted from the whole promoter. Results obtained with 16 features compared to 48 features indicates that a global signal is much more powerful than a local signal.

Annotation of same $section1$ and $section3$ of E.coli obtained with features extracted from interactions derived by Mandel et al. is done and the results of the annotation for $section1$ is shown in Figure 5. Similar results are predicted by these features also. But, the stretch of promoters is about 15-25 only. None of these results are predicting as well as 3-grams [6].

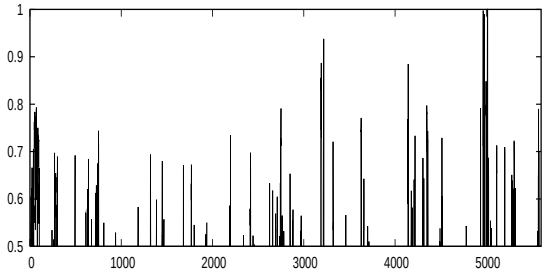

Figure 5: $section1$ of E.coli tested with 16 bi-gram features extracted from interactions proposed by Mandel et al. X-axis has a moving window of size 80 bp and y-axis shows the output of the neural network for each window.

Moreover, frequency analysis of the binding sites extracted using EIIP and Mandel values, indicates a marked bias towards certain bases in positions -35 and -31 and also -13, -12, -11 and -10. Table 5 and Table 6 represent the frequency of occurrence each base pair at each position in -35 binding site. The consensus at the -35 binding site of each position using EIIP gives a closer similarity to the general consensus TTGACA observed in literature. Since EIIP values for T and C are close, that could explain some distribution between T and C at -35 position and -31. In case of interactions obtained through Mandel's values, since, Glutamine favours A in comparison to the others, we could observe, a bias towards A. So is Arginine at position -31 which favours G.

The annotation results of these sections of E.coli compare with that of results obtained using 3-grams in the earlier work [6]. The distinct aspect of this work is the identification of the binding sites through the interactions between RNA polymerase and the binding sites of the promoter. If the binding sites were not identified correctly, the resulting bi-grams around the windows would not lead to the correct identification of the promoters. This was verified through experiments where the binding sites were incorrectly identified and the resulting classification accuracy of promoters was down to 45%. Only in the case of correct identification, we get to identify binding sites correctly hence can identify promoters much better.

Table 5: Frequency of occurrence of bases in -35 binding site for promoters using Mandel et al. interaction values

	-35	-34	-33	-32	-31	-30
A	**0.503**	0.282	0.298	0.230	0.018	0.242
T	0.381	0.285	0.367	0.430	0.228	0.317
G	0.113	0.175	0.089	0.094	**0.753**	0.145
C	0.0014	0.257	0.243	0.245	0.000	0.296

Table 6: Frequency of occurrence of bases in -35 binding site for promoters using EIIP encoding forinteraction

	-35	-34	-33	-32	-31	-30
A	0.051	0.291	0.260	0.341	0.052	0.294
T	0.412	0.375	0.288	0.323	0.374	0.269
G	0.000	0.224	0.309	0.224	0.000	0.291
C	**0.537**	0.109	0.142	0.112	**0.574**	0.145

5. Conclusions

Promoter recognition is attempted using the interactions between RNA polymerase and promoter. Experiments with similarity and cross correlation between RNA polymerase and promoter are tried. Experiments used to obtain similarity and cross-correlation using EIIP values show that a global signal (Figure 4) is rather more effective than a local signal (Figure 3). Eventhough the test data results indicate a higher sensitivity value, generalization capability of the 16 features is better than 48 features. The results also point to the fact that similarity measure between the signal is more efficient in promoter recognition. Interactions derived using amino acid-base pairs are not as powerful as the signal derived using EIIP values. The analysis of frequency distribution of bases in the binding sites shows that EIIP values have a distribution closer to the predicted consensus sequence compared to amino acid-base pair interactions. Additivity of interactions is assumed in these cases. Whether there is a stable conformation possible, with a lower inetraction value is to be investigated further. And also addition of more interactions to the set will increase the accuracy much further. A committee machine using these different features can be designed to annotate a segment as a promoter or a non-promoter based on voting. In addition the same sections are used for annotation with GLIMMER and Genemark packages which annotate the coding regions in the given DNA segment. Most of our promoters identified are occurring upstream of these coding regions giving credence to our annotation scheme.

Same arguments can be extended for eukaryotes in which

lot of transcription binding factors (TBP) bind to a promoter before a RNA polymerase is summoned. In this case, the interaction between TBPs and promoter can be modeled through the protein-promoter binding interactions and can be use to identify the promoters.

The main aim of the work is to prove the efficacy of the interaction between the RNA polymerase and the DNA in identifying the promoters in a whole genome. Eventhough the n-gram features are being used, it is very important to correctly identify the binding site regions through the interaction between DNA and RNA polymerase to get good accuracies. Hence, the main assumption that the interaction between DNA and RNA polymerase is proven to be very useful in promoter identification.

Acknowledgements

I would like to acknowledge the help of my student Mr.Naresh in the survey of literature and the financial assistance provided by my university, University of Hyderabad.

References

[1] V. Bajic, A. Chong, S. Seah, V. Brusic, " An Intelligent System for Vertebrate Promoter Recognition," *IEEE Intelligent Systems*, pp. 64-70 , 2002.

[2] Q. Li, H. Lina, " The recognition and prediction of σ70 promoters in Escherichia coli K-12," *Journal of Theoretical Biology*, vol. 242, pp. 135-141 , 2006.

[3] Y. Huang, C. Wang, " Integration of knowledge discovery and artificial intelligence approaches for promoter recognition in DNA sequences," *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)* , 2005.

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

[5] H. Ji, D. Xinbin, Z. Xuechun, " A systematic computational approach for transcription factor target gene prediction," *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology CIBCB '06*, 2006, p1.

[6] T. Sobha Rani, S. Bapi Raju, " Analysis of n-gram based promoter recognition methods and application to whole genome promoter prediction," *In Silico Biology* , vol. 9, pp. s1-s16, 2009.

[7] I. Deyneko, E. Alexander, B. Helmut, G. Kauer, " Signal-theoretical DNA similarity measure revealing unexpected similarities of E. coli promoters," *In Silico Biology* , vol. 5, online 2005.

[8] K. Aditi, B. Manju, "A novel method for prokaryotic promoter prediction based on DNA stability," *BMC Bioinformatics* , vol. 6, doi: 10.1186/1471-2105-6-1 , 2005.

[9] H. Wang, C. Benham, " Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress," *BMC Bioinformatics* , vol. 7, doi: 10.1186/1471–2105–7–248 , 2006.

[10] T. Abeel, P.R. Yvan Saeys, Y.V. de Peer, " ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles," *Bioinformatics* , vol. 24, pp. i24-i31, 2008.

[11] J. Ponomarenko, M. Ponomarenko, A. Frolov, D. Vorobyev, G. Overton, N. Kolchanov, " Conformational and physicochemical DNA features specific for transcription factor binding sites," *Bioinformatics* , vol. 15, pp. 654-668, 1999.

[12] L. Gordon, A.Y. Chervonenkis, A.J. Gammerman, I.A. Shahmurradov, V. Solovyev, " Sequence alignment kernel for recognition of promoter regions," *Bioinformatics* , vol. 19, pp. 1964-1971 , 2003.

[13] Q. Ma, J.T.L. Wang, D. Shasha, C.H. Wu, " DNA sequence classification via an expectation maximization algorithm and neural networks: a case study," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, Special Issue on Knowledge Management*," vol. 31, pp. 468-475 , 2001.

[14] A. Pedersen, P. Baldi, Y. Chauvinb, S. Brunak, " The biology of eukaryotic promoter prediction - a review," *Computers & Chemistry* , vol. 23, pp. 191-207 , 1999.

[15] T. Matsuda, H. Motoda, T. Washio, " Graph based induction and its applications," *Advanced Engineering Informatics* , vol. 16, pp. 135-143 , 2002.

[16] M. Gribskov, R.R. Burgess, " Sigma factors from E. coli, B. subtilis, phase SPO1, and phage T4 are homologous proteins," *Nucleic Acids Res.* , vol. 14, pp. 6745-6763 , 1986.

[17] J.D. Helmann, M.J. Chamberlin, " Structure and function of bacterial sigma factors," *Ann. Rev. Biochem.* , vol. 57, pp. 839-872 , 1988.

[18] J.A. Jaehing, " Sigma factor relatives in eukaryotes," *Science* , vol. 253, pp. 859 , 1991.

[19] M.J. Zvelebil, G.J. Barton, W.R. Taylor, M.J.E. Sternberg, " Prediction of protein secondary structure and active sites using the alignment of homologous sequences," *J.Mol. Biol.* , vol. 195, pp. 957-961 , 1987.

[20] D.A. Siegel, J.C. Hu, W.A. Walter, C.A. Gross, " Altered promoter recognition by mutant forms of the sigma 70 subunit of Escherichia coli RNA polymerase," *J. Mol. Biol.* , vol. 206, pp. 591-603 , 1989.

[21] T.J. Kenney, K. York, P. Youngman, C.P.J. Moran, " Genetic evidence that RNA polymerase associated with A factor uses a sporulation-specific promoter in Bacillus subtilis," *Proc. Natl. Acad. Sci. USA* , vol. 86, pp. 9109-9113 , 1989.

[22] C. Waldburger, T. Gardella, R. Wong, M.M. Susskind, "Changes in conserved region 2 of Escherichia coli sigma 70 affecting promoter recognition," *J. Mol. Biol.* , vol. 215, pp. 267-276 , 1990.

[23] A. Malhotra, E. Severinova, S.A. Darst, " Crystal structure of a σ70 subunit fragment from E. coli RNA polymerase," *Cell* , vol. 87, pp. 127-136 , 1996.

[24] T.Gardella, T.Moyle, M.M.Susskind, " A mutant Escherichia coli sigma 70 subunit of RNA polymerase with altered promoter specificity," *J. Mol. Biol.* , vol. 206, pp. 579-590 , 1989.

[25] E.A. Campbell, O. Muzzin, M. Chlenov, J.L. Sun, C.A. Olson, O. Weinman, M.L. Trester-Zedlitz, S.A. Darst, " Structure of the Bacterial RNA Polymerase Promoter Specificity σ Subunit," *Molecular Cell* , vol. 9, pp. 527-539 , 2002.

[26] T. Sobha Rani ans S. Bapi Raju, "E.coli promoter recognition through wavelets," *Proceeding of BioComp'08 , 2008 International conference on Bioinformatics and Computational Biology)*, 2008, p. 256.

[27] H.T. Chafia, F. Qian, I. Cosic, " Protein sequence comparison based on the wavelet transform approach," *Protein Engineering* , vol. 15, pp. 193-203 , 2002.

[28] I. Cosic, *Macromolecular Bioactivity: Is It Resonant Interaction Between Macromolecules?-Theory and Applications*, *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 1101-1114 , 1994.

[29] Y. Mandel-Gutfreund, H. Margalit, " Quantitative parameters for amino acidâĂŞbase interaction: implications for prediction of proteinâĂŞDNA binding sites," *Nucleic Acids Research* , vol. 26, pp. 2306-2312 , 1998.

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

[31] NCBI Viewer
Available: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=1786181