Indian Journal of Agricultural Sciences **84** (1): 119–23, January 2014/Article

Development of a tool for computational prediction of σ70 promoters in *Pseudomonas* spp using SVM and HMM approaches

MERIN K ELDO[1], M K RAJESH[2], T P JAMSHINATH[3], N HEMALATHA[4], MURALI GOPAL[5] and GEORGE V THOMAS[6]

Bioinformatics Centre, Central Plantation Crops and Research Institute, Kasaragod, Kerala 671 124

Received: 10 October 2012; Revised accepted: 8 October 2013

ABSTRACT

Promoters are regions in DNA that play important role in the regulation of gene expression. The ability to locate promoters within a section of DNA is known to be a very difficult and important task in DNA analysis. Since experimental techniques to identify promoters are costly and time consuming, *in silico* methods offer an alternative. In this study, we have developed a tool for identification of σ70 promoters in the –10 and –35 regions of sequences from *Pseudomonas* spp. Promoters were predicted using both Support Vector Machine (SVM) and Hidden Markov Model (HMM) based approaches. SVM performed better when trained using RBF kernel with a cross-validation of 5 and a value of 0.03 for the gamma parameter. The module developed using SVM showed a sensitivity of 78% and a specificity of 80%. The programmes required to process the user input were written using Perl and HTML codes were used to create a user interface. The user interface accepts a query sequence and the processed result will be displayed in a new window. The tool named 'PROMIT' (PROMoter Identification Tool), was developed in the Windows platform, has a user friendly interface and works well for sequences from *Pseudomonas* spp.

Key words : HMM, Promoter, *Pseudomonas*, SVM, σ70,

Promoters are regions of DNA where gene transcription is initiated and identification of promoters is of prime importance in deciphering the regulation of genes. The bacterial core RNA polymerase complex, which consists of five subunits (α2, β, β′, and ω), is sufficient for transcription elongation and termination, but is unable to initiate transcription. Transcription initiation from promoter elements requires a sixth, dissociable subunit called σ factor, which reversibly associates with the core RNA polymerase complex to form a holoenzyme (Borukhov and Nudler 2003). The holoenzyme, in turn, binds to its cognate promoters to initiate transcription (Browning and Busby 2004).

The vast majority of σ factors belong to the σ70 family, responsible for most of the housekeeping transcriptional activity (Boruknov and Nudler 2003). Members of the σ70 family direct RNA polymerase to specific promoter elements that are usually 5-6 base-pairs (bp) in length and are centered 10 and 35 bp upstream (positions –10 and –35) of the transcription initiation site. They also function in the melting of promoter DNA and the early stages of elongation of transcripts. The binding site for the σ70 family of promoters is defined by two consensus hexamers, TTGACA and TATAAT, located at approximately –35 and –10, respectively relative to the transcript start site. Prediction of these σ factor binding sites is an important first step in characterizing regulatory regions of DNA (Towsey *et al.* 2007a).

Various experimental techniques have been employed for promoter prediction, which, in spite of their accuracy, are time-consuming and expensive and therefore computational methods are required to facilitate faster processing of data. The development of promoter databases have permitted the employment of bioinformatic tools to predict the location of promoter regions based on either the homology to the consensus sequence or to a reference list of promoters (Polate and Gunes 2007). The most common and classic *in silico* approach for promoter prediction involves the development of algorithms which use position-weight matrices (PWMs) and relies on the relative conservation of –35 and –10 hexamers (Bucher 1990, Gordon *et al*. 2006). Other algorithms based on probability distributions relative to the transcript start site (Down and Hubbard 2002, Weller and Recknagel 1994) and Markov Models (MM) (Ohler *et al*.

[1]Research Scholar (e mail: mariaeldo@yahoo.co.in), [2]Senior Scientist (e mail: mkraju_cpcri@yahoo.com), [3]Project Assistant (e mail: unni.jamshi@gmail.com), [4]Professor (e mail: hemasree71@gmail.com), Bioinformatics Centre, AIMIT, St. Aloysius College, Mangalore, Karnataka 575 002, [5]Principal Scientist (e mail: mgcpcri@yahoo.co.in), [6]Director (e mail: georgevthomas@yahoo.com)

1999, Audic and Claverie 1997) has also been developed. Currently, machine learning approaches, especially Support Vector Machines (SVM) (Gordon and Towsey 2005), and Neural Networks (NN) (Demeler and Zhou 1991, Rani *et al*. 2007, de Avila *et al*. 2011) are being utilized for promoter recognition and prediction. A majority of the above algorithms have been developed for identification of promoters in *E. coli* and *Bacillus subtilis* (Maetschke *et al*. 2006, Towsey *et al.* 2007b).

The objectives of this study were (i) to develop an SVM-based tool that would determine the position and sequence of σ70 promoters in *Pseudomonas* spp, (ii) To develop an HMM model to determine the position and sequence of σ70 promoters in *Pseudomonas* spp and (iii) To compare and validate the result generated by the SVM-based tool with the result generated by the HMM model.

MATERIALS AND METHODS

The datasets for training and testing and creation of SVM and HMM models were obtained from the NCBI database in February 2011. The positive data consisted of 174 *Pseudomonas* promoter sequences derived from the experimental work of Swingle *et al.* (2008), whereas the negative data consisted of 306 sequences representing the coding regions of various bacterial genes.

The support vector machine (SVM), which is a machine learning method, has been applied for many kinds of pattern recognition problems. The principle of the SVM is to transform the samples into a higher dimension space called 'Hilbert space'. A separating hyperplane is sought in this space called the optimal separating hyperplane in such a way as to maximize its distance from the closest training samples. SVM is a supervised machine learning technology founded theoretically on Statistical Learning Theory (Vapnik 1995). *SVM*^{light} software version 6.02 [http://download.joachims.org/svm_light/current/svm_light_windows.zip] was used for data classification using SVM. The software enables users to define a number of parameters as well as inbuilt kernel functions such as linear kernel, radial basis function and polynomial kernel of a given degree.

The methodology followed for computational prediction of σ70 promoters in *Pseudomonas* spp is summarized in Fig 1. For prediction of promoters using SVM, the sequence data was first converted into the input format of the *SVM*^{light} software. It was decided to use four-mer frequency as the input because only tetramers are found to be conserved in the –35 and –10 canonical hexamers of most of the bacterial promoters (Gordon *et al*. 2006). Supplementary Table 5 shows the four-mers considered for frequency calculation. *SVM_learn* and *SVM_classify* modules were then used for training and testing SVM. Cross-validation tests were also carried out. The testing set was used to evaluate the performance of SVM. A confusion matrix with True Positive, True Negative, False Positive, and False Negative was created to determine the performance of SVM on testing set. Sensitivity (S_n) and specificity (S_p) were calculated using the equation given below:

$$S_n = TP/(TP + FN) \quad S_p = TP/(TP + FP)$$

Sensitivity determines the accuracy with which the positive datasets are classified as the positive class and specificity determines the accuracy with which negative datasets are classified as negative. Thus SVM parameters should be optimized in a way that ensures high sensitivity and specificity. To optimize the SVM output, training was repeated several times with different parameters (linear, polynomial and RBF kernels) and after each training phase, the sensitivity and specificity were calculated. Different cross-validation values (Chou and Zhang 1995) were used while training.

The same sets of sequences used for prediction using SVM were used to generate the HMM models– one for recognizing promoter sequences and the other for non-promoter sequences. Perl programmes were written to calculate the initial probability, transition probability and emission probability.

RESULTS AND DISCUSSION

The SVM^{light} software was trained using different sets of sequences and different SVM parameters. For each set of sequences and parameters, the accuracy of classification by SVM was determined by calculating both sensitivity and specificity. On observing the performance of SVM in each of these cases, it was concluded that for the sequences used in the study, the machine learning approach works well when trained using the RBF kernel. In general, the RBF kernel is a reasonable first choice because this kernel non-linearly maps samples into higher dimensional space. Unlike the linear kernel, the RBF kernel can handle the case when the relationship between class labels and the attributes is non linear (Hsu and Lin 2002).

Cross validation was another statistical measure which was used to estimate the predictive accuracy of classifier on data. Here, each sample data was singled out in turn as the test set and the remaining sequences were used as the training set to distinguish if the sequences in the test set are promoters or not. Though different cross validation values were used, best results were obtained when SVM was trained with a cross validation value of 5. Using the RBF kernel and a cross validation value of 5, the tool showed a sensitivity of 78% and a specificity of 80%. The sensitivity and specificity values obtained here are more when compared to identification of σ70 promoters reported earlier mainly utilizing neural network approaches (Demeler and Zhou 1991, Rani *et al*. 2007, de Avila *et al.* 2011)

To optimize the prediction performance, the SVM was trained using different combinations of positive and negative datasets, while maintaining the same set of parameters and

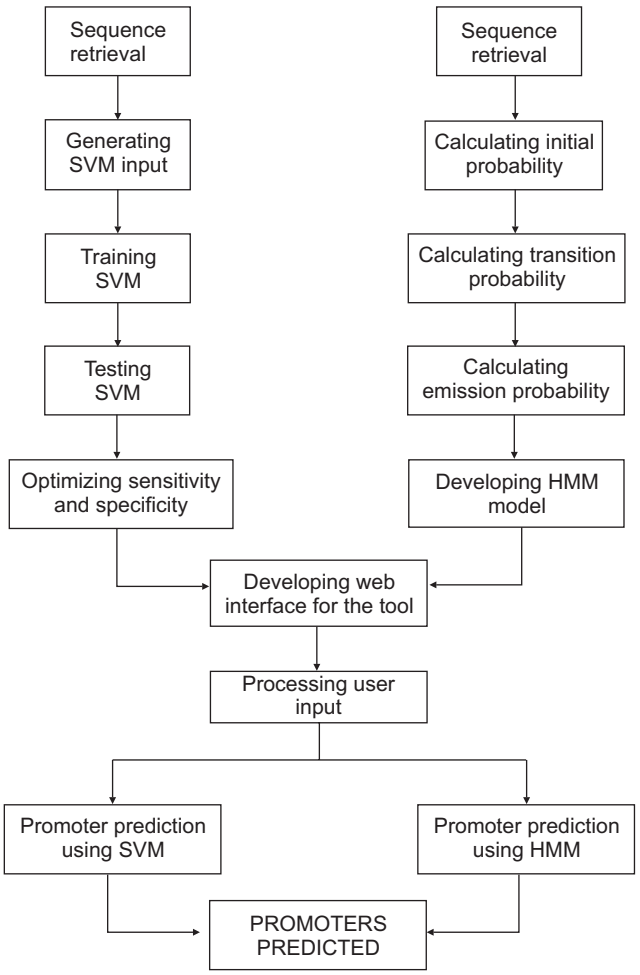

Fig 1 The flow chart of PROMIT approach

keeping the kernel constant. No significant change was observed in the sensitivity and specificity, thus justifying the use of the RBF kernel and cross validation value of 5. The tool was tested using non-promoter sequences like the coding sequences which were correctly classified as non-promoters.

Promoter prediction was also carried out using another computational approach, HMM. The first model (HMM model 1) contained the initial, transition and emission probabilities calculated from the promoter sequences, while the second model (HMM model 2) contained the probabilities calculated from the negative sequences. The initial transition and emission probabilities were calculated from the initial sequences datasets (Tables 1 and 2). The comparison of the promoter sequences predicted by the HMM approach with already characterized promoter sequences clearly indicates the improved performance of Hidden Markov Models in promoter prediction. Similar to the prediction task performed by SVM, HMMs also predict $\sigma 70$ promoters from the sequence submitted by the user.

Another objective of the work was to compare and validate the promoters predicted by SVM and HMM approach. With a sensitivity of 78% and a specificity of 80%, the results showed that several promoters predicted by the SVM tool matched with the promoters already characterized in the species. However, there were differences in the predictions made by the two approaches. The difference in promoters predicted by the two approaches can be attributed to the fact that the two computational techniques used are based on different theories and both differ in their basic working. Studies are required in the direction of improving the positional ability of the approaches, datasets utilized by programs etc. to further improve the accuracy of predictions.

The tool was developed using Perl programming language. When a user submits a query sequence, the tool first identifies probable promoter sequences (40 bp preceding the ATG/TTG/GTG codons). The sequences were then classified using the SVM model and the two HMM models developed. The user friendly web interface of PROMIT (PROMoter Identification Tool) was developed using HTML and Javascript. While all the programs required to process

Table 1 Initial matrix for promoter sequence

Nucleotide	Probability
T	0.85
A	0.07
G	0.04
C	0.05

Table 2 Transition matrix for promoter sequences

Nucleotide → ↓	A	T	G	C
T	170	333	67	222
A	286	164	83	94
G	100	124	49	121
C	108	120	192	130

Table 3 Initial matrix for non-promoter sequences

Nucleotide	Probability
T	0.29
A	0.23
G	0.25
C	0.23

Table 4 Transition matrix for non promoter sequences

Nucleotide → ↓	A	T	G	C
T	0.26	0.13	0.34	0.26
A	0.28	0.25	0.22	0.24
G	0.21	0.23	0.29	0.27
C	0.23	0.21	0.30	0.25

Table 5 Four-mers considered for frequency calculation using SVM

AAAA	AAAT	AAAG	AAAC	AATA	AATT	AATG	AATC	AAGA	AAGT
AGCA	AACA	AACT	AACG	AACC	ATAA	ATAT	ATAG	ATAC	ATTA
ATTT	ATTG	ATTC	ATGA	ATGT	ATGG	ATGC	ATCA	ATCT	ATCG
ATCC	AGAA	AGAT	AGAG	AGAC	AGTA	AGTT	AGTG	AGTC	AGGA
AGGT	AGGG	AGGC	AGCA	AGCT	AGCG	AGCC	ACAA	ACAT	ACAG
ACAC	ACTA	ACTT	ACTG	ACTC	ACGA	ACGT	ACGG	ACGC	ACCA
ACCT	ACCG	ACCC	TAAA	TAAT	TAAG	TAAC	TATA	TATT	TATG
TATC	TAGA	TAGT	TAGG	TAGC	TACA	TACT	TACG	TACC	TTAA
TTAT	TTAG	TTAC	TTTA	TTTT	TTTG	TTTC	TTGA	TTGT	TTGG
TTGC	TTCA	TTCT	TTCG	TTCC	TGAA	TGAT	TGAG	TGAC	TGTA
TGTT	TGTG	TGTC	TGGA	TGGT	TGGG	TGGC	TGCA	TGCT	TGCG
TGCC	TCAA	TCAT	TCAG	TCAC	TCTA	TCTT	TCTG	TCTC	TCGA
TCGT	TCGG	TCGC	TCCA	TCCT	TCCG	TCCC	GAAA	GAAT	GAAG
GAAC	GATA	GATT	GATG	GATC	GAGA	GAGT	GAGG	GAGC	GACA
GACT	GACG	GACC	GTAA	GTAT	GTAG	GTAC	GTTA	GTTT	GTTG
GTTC	GTGA	GTGT	GTGG	GTGC	GTCA	GTCT	GTCG	GTCC	GGAA
GGAT	GGAG	GGAC	GGTA	GGTT	GGTG	GGTC	GGGA	GGGT	GGGG
GGGC	GGCA	GGCT	GGCG	GGCC	GCAA	GCAT	GCAG	GCAC	GCTA
GCTT	GCTG	GCTC	GCGA	GCGT	GCGG	GCGC	GCCA	GCCT	GCCG
GCCC	CAAA	CAAT	CAAG	CAAC	CATA	CATT	CATG	CATC	CAGA
CAGT	CAGG	CAGC	CACA	CACT	CACG	CACC	CTAA	CTAT	CTAG
CTAC	CTTA	CTTT	CTTG	CTTC	CTGA	CTGT	CTGG	CTGC	CTCA
CTCT	CTCG	CTCC	CGAA	CGAT	CGAG	CGAC	CGTA	CGTT	CGTG
CGTC	CGGA	CGGT	CGGG	CGGC	CGCA	CGCT	CGCG	CGCC	CCAA
CCAT	CCAG	CCAC	CCTA	CCTT	CCTG	CCTC	CCGA	CCGT	CCGG

the user input were written using Perl, HTML codes were used to create a user interface. The web interface allows users to submit a query and the output will be displayed in a new window. The interface was developed in a user friendly manner and contains web page that serves different purposes, which are highlighted below. The home page gives a one line description about what the tool does. Users can either paste a FASTA sequence in the text area or can upload a valid FASTA file. The tool accepts only one sequence at a time. Alert messages are displayed if the user submits a blank entry, sequence in a format other than FASTA or a sequence with characters other than ATGC. PROMIT's homepage provides five links. Each of these links has one or more sub links, displayed in the form of a dropdown list. The content of each link or sub link will be displayed in a new web page. The options include: 'Home' (this option takes the user from any other page to the homepage) and '*About'* (this link provides a brief description about the various building blocks of the tool). Sub-links include '*Pseudomonas'* (provides a detailed description about the general characteristics of this group of organisms. The section also throws light into the various species coming under this class, their uses etc.), '*Promoter'* (the function, structure and other information about promoters are detailed in this section), '*SVM'* (this section deals with the basics of the machine learning approach SVM. A short history, brief working and the basic mathematics behind SVM is displayed here) and 'HMM' (The basic idea behind Hidden Markov Model is given under the sub link 'HMM'. The section also deals the basics and working of HMM), 'HMM TOOL' (This link provides the tool that predicts promoters using HMM. Users can either paste a sequence or upload a file. On clicking the 'Process' button, the output will be displayed in a new window), 'Help' (This link provides access to the web page with a set of instructions which the user may follow to get an idea on how to use the tool) and 'Main Links' (Three sub links are provided by this option. Each of these sub links point to the home pages of: CPCRI, Kasaragod and Bioinformatics Centre and Library, CPCRI)

The output generated by the tool is displayed in a new window. If the tool detects promoters in the sequence submitted, then the result page displays the 'Sequence accession number', 'Number of promoters predicted', 'Serial number', 'Start position of promoter', 'End position of promoter' and also provides the 'Promoter sequence'. If the sequence does not contain any promoters, then the result is displayed with the accession number of the query sequence and a message that no promoters were detected.

For very large sequences, like the whole genome sequence, the tool (both SVM and HMM based tool) recommends users to run the promoter prediction process offline. For this purpose, the tool provides a zipped file,

which contains the Perl programs that process user's query and a readme file which contains a set of instructions to run the tool offline. The zipped file is downloadable. On running the Perl program in command prompt, the promoters predicted will be written into a text file and stored in the user's computer.

To conclude, PROMIT introduces a method for promoter prediction that enables convenient description of both SVM and HMM based promoters. Promoter sequences and coding sequences from the *Pseudomonas* spp were used to train the SVM and HMM models. The trained models were then used for the prediction of promoters in a user submitted sequence. The successful prediction of promoters shows that the method followed has significant merit as an approach for prediction of promoters in other organisms sharing a close evolutionary relationship with the *Pseudomonas* species. There are still several directions in which the task of computational prediction of promoters can progress. This, in turn, throws light into the fact that in the near future the importance of promoter prediction techniques will only increase. As more and more genomes are sequenced, there would be an increasing demand for more accurate computational techniques to extract knowledge from these vast amounts of data.

ACKNOWLEDGMENT

This work was supported by a grant from Department of Biotechnology (BTISnet) and Department of Information Technology (ABPC), New Delhi, India.

REFERENCES

Audic S and Claverie J M. 1997. The significance of digital gene expression profiles. *Genome Research* **7**: 986–95.

Borukhov S and Nudler E. 2003. RNA polymerase holoenzyme: structure, function and biological implications. *Current Opinion in Microbiology* **6**: 93–100.

Browning D F and Busby S J W. 2004. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology* **2**: 57–65.

Bucher P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology* **212**: 563–78.

Chou K C and Zhang C T. 1995. Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* **30**: 275–349.

de Avila E S S, Gerhardt G J and Echeverrigaray S. 2001. Rules extraction from neural networks applied to the prediction and recognition of prokaryotic promoters. *Genetics and Molecular Biology* **34**: 353–60.

Demeler B and Zhou G. 1991. Neural network optimization for *E. coli* promoter prediction. *Nucleic Acids Research* **19**: 1 593–99.

Down T A and Hubbard T J. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Research* **12:** 458–61.

Gordon J and Towsey M. 2005. SVM based prediction of bacterial transcription start sites. *Proceedings of 6th International Conference on Intelligent Data Engineering and Automated Learning*, July 2005, Brisbane, pp 448.

Gordon J, Towsey M, Hogan J, Mathews S and Timms P. 2006. Improved prediction of bacterial transcription start sites. *Bioinformatics* **22**:142–8.

Hsu C-W and Lin C-J. 2002. A simple decomposition method for support vector machines. *Machine Learning* **46**: 291–314.

Maetschke S R, Towsey M and Hogan J M. 2006. Bacterial promoter modelling and prediction for *E. coli* and *B. subtilis* with Beagle. *Workshop on Intelligent Systems for Bioinformatics (WISB–2006)*, 4th December 2006, Hobart, Tasmania.

Ohler U, Harbeck S, Niemann H, Noth E and Reese M G. 1999. Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* **15**: 362–9.

Polate K and Gunes S. 2007. A novel approach to estimation of *E. coli* promoter gene sequences: Combining feature selection and least square support vector machine (FS_LSSVN). *Applied Mathematics and Computation* **190**: 1 574–82.

Rani T S, Bhavani S D and Bapi R S. 2007. Analysis of *E. coli* promoter recognition problem in dinucleotide feature space. *Bioinformatics* **23**: 582–88.

Swingle B, Thete D, Moll M, Myers C R, Schneider D J and Cartinhour S. 2008. Characterization of the PvdS-regulated promoter motif in *Pseudomonas syringae* pv. *tomato* DC3000 reveals regulon members and insights regarding PvdS function in other pseudomonads. *Molecular Microbiology* **68**: 871–89.

Towsey M, Hogan J M, Mathews S and Timms P. 2007a. The *in silico* prediction of promoters in bacterial genomes. *Genome Informatics* **19**: 178–89.

Towsey M, Timms P, Hogan J M and Mathews S. 2007b. The cross-species prediction of bacterial promoters using a support vector machine. *Computational Biology and Chemistry* **32**: 359–66.

Vapnik V. 1995. *The Nature of Statistical Learning Theory*. pp 188. A Springer-Verlag publication, New York.

Weller K and Recknagel R D. 1994. Promoter strength prediction based on occurrence frequencies of consensus patterns. *Journal of Theoretical Biology* **171**: 355–9.