PCA-HPR: A principle component analysis model for human promoter recognition

Xiaomeng Li[1, 2, 3, *], Jia Zeng[1] and Hong Yan[1, 2]

[1]Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong; [2]School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia; [3]School of Astronautics, Harbin Institute of Technology, Harbin, China; Xiaomeng Li* - E-mail: sgusico@hotmail.com; * Corresponding author

received January 29, 2008; revised May 06, 2008; accepted May 09, 2008; published June 19, 2008

Abstract:

We describe a promoter recognition method named PCA-HPR to locate eukaryotic promoter regions and predict transcription start sites (TSSs). We computed codon (3-mer) and pentamer (5-mer) frequencies and created codon and pentamer frequency feature matrices to extract informative and discriminative features for effective classification. Principal component analysis (PCA) is applied to the feature matrices and a subset of principal components (PCs) are selected for classification. Our system uses three neural network classifiers to distinguish promoters versus exons, promoters versus introns, and promoters versus 3' un-translated region (3'UTR). We compared PCA-HPR with three well-known existing promoter prediction systems such as DragonGSF, Eponine and FirstEF. Validation shows that PCA-HPR achieves the best performance with three test sets for all the four predictive systems.

Keywords: promoter recognition; sequence feature; CpG islands; transcription start sites; principal component analysis

Background:

Eukaryotic promoter prediction plays a very important role in the study of gene regulation. Improvements are needed despite the availability of a number of promoter prediction algorithms. There is a need to increase true positive predictions and at the same time reduce false positive predictions. The most important issue is the selection of appropriate features for developing prediction systems.

Available promoter prediction systems use two types of features for classification namely, context features like n-mers, and signal features such as TATA-box, CCAAT-box, and CpG islands. Among the favorable promoter prediction programs, Eponine **[1]** builds a PWM to detect TATA-box and G+C enrichment regions as promoter-like regions; FirstEF **[2]** uses CpG-related and non-CpG related first exons as signal features; PromoterInspector **[3]** uses IUPAC words with wildcards as context features. Good experiment results are achieved by integrating these two types of features. DPF **[4]** applies a separate module on G+C rich and G+C poor regions, and selects 256 pentamers to generate a PWM for prediction. Furthermore, DragonGSF **[5, 6]** adds the CpG-island feature to DPF.

However, the performance of existing methods is still not satisfactory. There is a common problem in these prediction systems and they select limited number of features for classification. So, they ignore information in abandoned features and the interaction of selected features. Feature vectors need to be rebuilt to include more information for classification to achieve better prediction results.

Here, we describe a method named PCA-HPR to predict the location of the TSSs with best performance. Principle Component Analysis (PCA) is applied to the context feature selection from feature matrices. The level of information loss was controlled by choosing a certain number of principal components (PCs). PCA-HPR projects original features extracted from training sequences to a new feature space constructed by PCs instead of choosing specific pentamers, (e.g., CGGCG, GCGCG) which are used in PromoterExplorer **[7]**. Resulting feature vectors are then sent to artificial neural networks (ANNs) for training. The concept of CpG islands (genomic regions that contain a high frequency of CG di-nucleotides) is also used as an enhancing signal. The final prediction is performed by a data processing module which combines the output from three classifiers and a CpG island module. The positive predictive value (PPV) and sensitivity (SN) of the model are shown to be higher than existing methods.

Methodology:

The training set in this experiment is divided into several subsets of promoters, introns, exons and 3′UTR sequences. Promoter sequences are extracted from two public databases. One is the Eukaryotic Promoter Database (EPD), release 86 **[8]**, which contains 1871 human promoter sequences. The other is the Database of Transcription Start Sites (DBTSS), version 5.2.0 **[9]**, which includes 30,964 human promoter sequences and 15,531 forward strand promoter sequences. We used forward strand promoter sequences are in our experiment. Human exon and intron sequences are extracted from the exon-intron database **[10]**, and the first exons are not included in the exon training set. Human 3'UTR sequences are from the UTR database **[11]**.

We randomly selected 1000 promoter sequences from EPD, and 7000 promoter sequences from DBTSS to form the promoter training set. We then extracted 250bp upstream to 50bp downstream relative to the TSS of promoter sequences. In the non-promoter dataset, we selected sequences longer than 1200bp (compared to sequence of length 1200bp in EPD). We arranged the selected sequences into 300bp each and choose 10,000 sequences each from Exon, Intron and UTR databases.

CpG islands, which exist in 60% of mammalian promoters **[12]**, are regarded as one of the most important signal features in promoter recognition. Methods such as CpGProD **[13]**, DPF **[4]**, DragonGSF **[5, 6]**, FirstEF **[2]** and PromoterExplorer **[7]** embed this signal feature in their prediction system. In our method, two features are used to identify if a given sequence (>200bp) is CpG islands related: GC percentage (GCp) and Observed/expected CpG ratio (o/e). These are calculated with equations (1-5) (see supplementary material). If GCp > 0.5, and o/e > 0.6, then the sequence is considered CpG islands related, otherwise it is non-CpG islands related **[14]**.

A DNA sequence contains four types of nucleotides: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). With different combinations, there are 4^3=64 codons and 4^5=1024 pentamers in promoter and non-promoter datasets. Pentamers are widely selected as context features in many promoter prediction models **[4, 7]**, as they keep good balance between search efficiency and discriminability. In PCA-HPR, we also extract frequency matrices of pentamers. Additionally, we evaluate the contribution of pentamers and codons to the separation between the promoter and non-promoter sequences using the relevance function (6) (see supplementary material). Then the statistical relevance values of 64 codons and 1024 pentamers are ranked, and the analysis result is shown in Figure 1. A larger relevance function value ***R*** represents a higher discriminative ability of the feature. Among the 100 features with highest ***R*** value, we found 30% to 40% of them are codons. Since the total number of codons is much less than the total number of pentamers, we conclude that codons have a good discriminative ability. So using codons together with pentamers as context features will improve the classification performance.

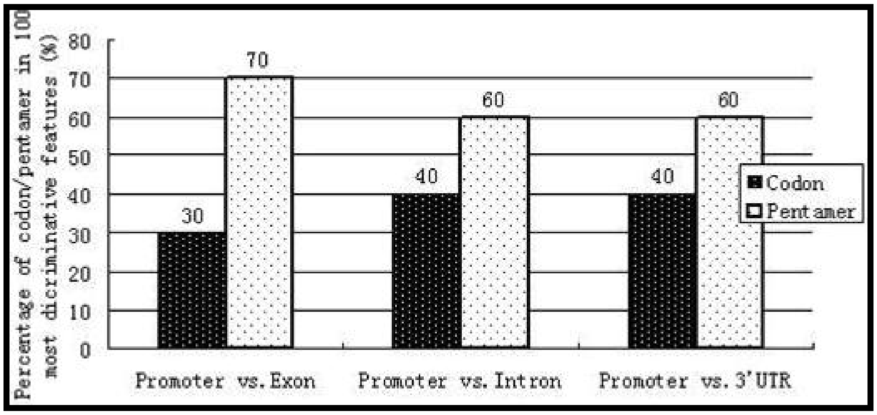

Figure 1: Codon/pentamer percentage in top 100 discriminative features. Statistics is based on three datasets: Promoter versus Exon, Promoter versus Intron, and Promoter versus 3'UTR.

To construct codon and pentamer combined frequency matrices, we first count overlapping codon and pentamer frequencies of fixed length sequences in promoter and non-promoter training sets. A 64×n matrix and a 1024×n matrix are built for each training set, where n represents the number of sequences of the training set. Next, we applied normalization and combined two matrices using Equations (7) to (9) for each dataset. Four resulting feature matrices are constructed from promoter, exon, intron and 3'UTR training datasets. Finally, three pairs of matrices: promoter versus exon, promoter versus intron and promoter versus 3'UTR are built from the four feature matrices for further processing.

Embedding all the codon and pentamer features in our system is not practical due to system efficiency. Moreover, redundant information provides noise and directly influences the prediction result of the system. In order to obtain more discriminative information in a relatively low-dimensional feature space, PCA is used in the context feature selection. PCA is an efficient method of reducing dimension of a dataset while retaining useful information, which accounts for most information of the original dataset. We apply PCA to the three pairs of codon and pentamer combined matrices. The idea of PCA in our experiment is to diagonalize the covariance matrix of the feature matrix. The diagonal elements are the variance data type, and the off-diagonal elements are the covariance data type in the covariance matrix. Here, large values of variance elements represent the signals that are of interest, and small ones represent noise. Large values of covariance elements correspond to high redundancy and small ones correspond to low redundancy. Therefore, the ideal solution is to find a matrix to diagonalize the covariance matrix by linear transformation, making the off-diagonal elements of the matrix zero. Through deduction (equations (10), (11) shown in supplementary material), we can see the matrix constructed with eigenvectors of the covariance matrix is the one that can diagonalize the covariance matrix. Larger eigen values of covariance matrix associate with higher levels of energy and the corresponding eigenvectors are the principal components (PCs) needed. The minimum number of vectors needed in our method is calculated according to Equation (12) (shown in supplementary material). Finally, the first six principal components are selected to form a new feature space in each of the three promoter and non-promoter matrix pairs.

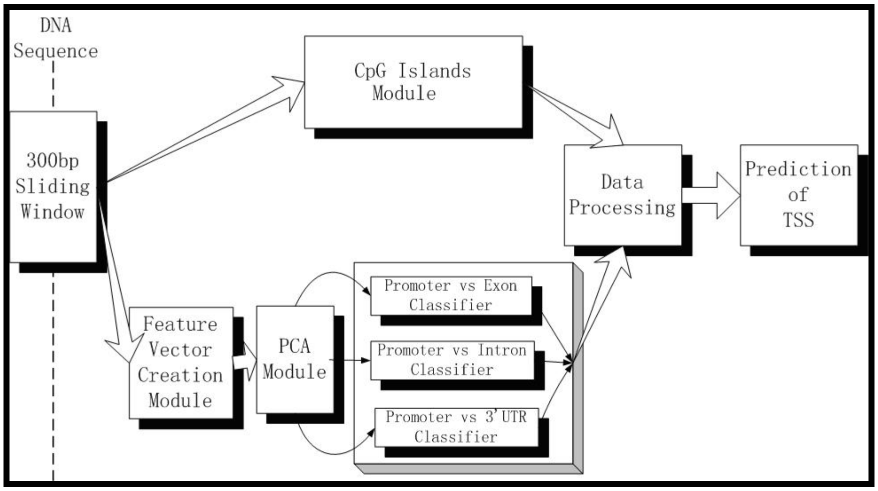
DNA Sequence
CpG Islands Module
300bp Sliding Window
Data Processing
Prediction of TSS
Promoter vs Exon Classifier
Feature Vector Creation Module
PCA Module
Promoter vs Intron Classifier
Promoter vs 3'UTR Classifier

Figure 2: Overall structure of our promoter recognition system. The 300bp sliding window moves 20bp at each step. Each sequence segment inside the sliding window is sent to the Feature Vector Creation Module and the CpG Islands Module. The Feature Vector Creation Model extracts codon and pentamer combined vector from the input sequence. The PCA module contains three spaces formed by PCs: the promoter-exon space, the promoter-intron space, and the promoter-3'UTR space. The vector generated by Feature Vector Creation Module is projected to the three spaces in the PCA model, and then sent to the three corresponding Classifiers. The three scores from the classifiers together with the score from CpG islands are processed in the Data Processing Model, and then the final prediction of TSSs is produced.

Traditional feature selection methods, such as DPF and PromoterExplorer, employ a certain number of pentamers, which is far less than the size of the original feature set. Although these pentamers are selected as the most discriminative features according to probability or distance functions, there is still massive information missed with the abandoned features. However, PCs selected by PCA contain information of most features in the datasets. Thus, they can best describe the characteristics of each dataset and improve classification. The PCA-HPR system is constructed with several sub-modules. Figure 1 shows the overall structure of the promoter recognition system. In the system, the CpG module gives a score for each input sequence segment: “1” for a CpG-island-related segment, and “0” for a non-CpG-island-related segment. Three classifiers in the model are built with ANNs, and each classifier is trained by 8000 promoter sequences and 10000 non-promoter sequences. The threshold of each classifier is set to 0.4, and if the outputs of two of the three classifiers are over the predefined threshold, the data processing module will sum the outputs of the three classifiers and the score from the CpG island module. If the sum is over 2.2, the data processing module will report the window as the potential promoter region. In the TSS prediction module, a promoter region is identified if the number of consecutive windows is more than 20, and the consecutive windows are defined here if the offset of two windows is less than 300bp. The predicted TSS is the location that contains the maximum score.

Discussion:

Three test sets are formed to evaluate the performance of PCA-HPR. Test set 1 consists of four human genomic sequences from GenBank with a total length of 0.95Mb and 14 known TSSs. The accession numbers of these sequences are L44140, D87675, AF017257 and AC002368. They are selected because most existing promoter prediction systems have tested them and the results are available for a fair comparison. Test set 2 uses the Chromosome 22 sequence and its annotation data **[15]**. The sequence has a total length of 34.75Mbp with 393 annotated TSSs. In test set 3, seven *Homo sapiens* chromosome 22 genomic sequences are extracted from GenBank with a total length of 11.56Mbp and 94 TSSs in the forward strands. The accession numbers of these sequences are NT_028395.3, NT_011519.19, NT_011521.4, NT_011523.11, NT_011525.7, NT_019197.5 and NT_011526.6. These sequences have different annotations than the one in test set 2, so the results are credible.

We selected three promoter systems, DragonGSF, Eponine and FirstEF to compare the performance on test set 1. A promoter region is counted as a true positive (TP) if TSS is located within the region, or if a region boundary is within 200bp 5' of such a TSS. Otherwise the predicted region is counted as a false positive (FP). The test results of Eponine and FirstEF are cited from the reference paper **[16]**. On test set 2, we adopt the same evaluation method as DragonGSF **[5]**: when one or more predictions fall in the region of [−2000, +2000] relative to a TSS, a TP is counted. All predictions which fall on the annotated part of the gene in the region [+2001, EndofTheGene] are counted as FP. Other predictions are not considered in counting TP and FP. Experimental results of DragonGSF, FirstEF and Eponine are obtained from **[5]**. We adopt the sensitivity (SN) and the positive predictive value (PPV) to evaluate the performance of these systems. The results and comparisons based on test 1 and test 2 are shown in Table 1 and Table 2 (tables in supplementary material).

In test 1, with the same number of true positives in comparison with existing methods, our method produces the smallest number of false positives. In test 2, although FirstEF achieves a higher SN than PCA-HPR, the PPV is just half of PCA-HPR. DragonGSF keeps a good balance between SN and PPV, while PCA-HPR produces better results. On test set 3, we compare PCA-HPR with DragonGSF because DragonGSF is the only online system which can accept relatively longer sequences among systems compared in the analysis. In order to get fair results for these sequences which are longer than 1,000,000bp (the limitation of a file in the DragonGSF web tool), we divided them into segments that are equal or less than 1,000,000bp each, before sending them to PCA-HPR and DragonGSF. Under the same evaluation criteria as the one in test set 2, PCA-HPR achieved a better result: the SN of PCA-HPR and DragonGSF are 53.2% and 46.8%, and PPV of the two systems are 72.4% and 63.8%, respectively. DragonGSF reports a good prediction performance on the whole human genome sequence, but it uses the TRANSFAC **[17]** database which includes binding site information only available for known promoters. Therefore, our system has the advantage in predicting unknown promoters.

Conclusion:

We have proposed a new system called PCA-HPR for promoter detection in DNA sequences. In this experiment, we focus on improving the feature selection process to achieve a better prediction performance. The majority of promoter prediction methods available now directly extract a limited number of context features from sequences. We use PCA to reduce the high dimensional feature matrices, and select PCs to form the new feature space. The promoter prediction method based on the rebuilding feature vectors is tested on three test datasets. The result of test sets 1 and 3 show that PCA-HPR can reduce false positive rate leading to a high PPV. Predictions on the genome sequence of chromosome 22 made by PCA-HPR are competitive in terms of SN and PPV. The comparison results indicate that the PCA algorithm performs effectively on feature selection, which is one of the most important tasks in human promoter recognition.

Acknowledgement:
This work is supported by the Hong Kong Research Grant Council (Project CityU122607).

References:

[01] T. A. Down *et al., Genome Research*, 12: 458 (2001) [PMID: 11875034]
[02] R. V. Davuluri *et al., Nature Genetics,* 29: 412 (2001) [PMID: 11726928]
[03] M. Scherf *et al., Journal of Molecular Biology,* 297: 599 (2000) [PMID: 10731414]
[04] V. B. Bajic *et al., Journal of Molecular Graphics and Modelling*, 21: 323 (2003) [PMID: 12543131]
[05] V. B. Bajic *et al., Genome Res.,* 13: 1923 (2003) [PMID: 12869582]
[06] V. B. Bajic *et al. Nature Biotechnology,* 22: 1467 (2004) [PMID: 15529174]
[07] X. Xie *et al., Bioinformatics,* 22: 2722 (2006) [PMID: 17000749]
[08] C. D. Schmid *et al., Nucleic Acids Research,* 34: 82 (2006) [PMID: 16381980]
[09] Y. Suzuki *et al., Nucleic Acids Research,* 30: 328 (2002) [PMID: 11752328]
[10] http://hsc.utoledo.edu/bioinfo/eid/index.html
[11] G. Pesole *et al., Nucleic Acids Research,* 30: 6 (2001) [PMID: 10592223]
[12] S. H. Cross *et al., Nucleic Acids Research,* 27: 2099 (1999) [PMID: 10219082]
[13] L. Ponger *et al., Bioinformatics,* 18: 631 (2002) [PMID: 12016061]
[14] M. Gardiner-Garden *et al., Journal of Molecular Biology,* 196: 261 (1987) [PMID: 3656447]
[15] http://www.sanger.ac.uk/HGP/Chr22
[16] S. Wu *et at., Physicl Review E,* 75: 041908 (2007) [PMID: 17500922]
[17] V. Matys *et al., Nucleic Acids Research*, 34: 108 (2006) [PMID: 16381825]

Edited by P. Kangueane
Citation: Li *et al.*, Bioinformation 2(9): 373-378 (2008)

Supplementary material

Equations:

$$GCp = P(C) + P(G) \qquad \rightarrow \qquad \mathbf{(1)}$$

$$o/e = \frac{P(CG)}{P(C) \times P(G)}, \qquad \rightarrow \qquad \mathbf{(2)}$$

$$P(C) = \frac{number\ of\ Cs}{length} \qquad \rightarrow \qquad \mathbf{(3)}$$

$$P(G) = \frac{number\ of\ Gs}{length} \qquad \rightarrow \qquad \mathbf{(4)}$$

$$P(CG) = \frac{number\ of\ CGs}{length} \qquad \rightarrow \qquad \mathbf{(5)}$$

$$R = \frac{(m_1 - m_0)^2 \times (\eta_1 - \eta_0)^2}{d_1^2 + d_0^2 + 1} \qquad \rightarrow \qquad \mathbf{(6)}$$

where m_1 and m_0 are the mean values of the occurrence number of the feature X in the promoter and non-promoter training sets. η_1 and η_0 are the percentage of promoters and non-promoters in which the feature X appears respectively. d_1 and d_0 are the standard deviation of the feature X in promoter and non-promoter training sets. In our work, the feature X denotes a pentamer or a codon.

$$a(i_1, j) = \frac{a_0(i_1, j)}{a_{\max}} \qquad i_1 = 1,2,...64 \quad j = 1,2,...n \qquad \rightarrow \qquad \mathbf{(7)}$$

$$b(i_2, j) = \frac{b_0(i_2, j)}{b_{\max}} \qquad i_2 = 1,2,...1024 \quad j = 1,2,...n \qquad \rightarrow \qquad \mathbf{(8)}$$

$$c_0(i_{3,} j) = \begin{bmatrix} a(i_1, j) \\ b(i_2, j) \end{bmatrix} \qquad i_3 = 1,2,...1088 \quad j = 1,2,...n \qquad \rightarrow \qquad \mathbf{(9)}$$

where $a_0(i_1,j)$, $a(i_1,j)$ and $b_0(i_2,j)$, $b(i_2,j)$ are codon and pentamer frequency matrix elements before and after normalization; a_{max} and b_{max} are maximum values of codon matrix and pentamer matrix, respectively. After normalization, the two matrices are integrated into one 1088×n feature matrix $c_{0,}$ and $c(i_3,j)$ represents the element in it.

(PCA in context feature selection)
Let us refer to C as a 1088×m feature matrix, where m is the total number of promoter and non-promoter samples in each pair. P is an orthonormal matrix, where $P^{-1}=P^T$.
Let Y=PC, so Y is the projection of C based on new space P.

$$\begin{aligned} C_Y &= \frac{1}{n-1} YY^T \\ &= \frac{1}{n-1}(PC)(PC)^T \qquad \rightarrow \qquad \mathbf{(10)} \\ &= \frac{1}{n-1} PCC^T P^T \\ &= \frac{1}{n-1} PAP^T \end{aligned}$$

where $A \equiv CC^T$.

As A is symmetric, we can find matrix E and D so that $A=EDE^T$, where D is a diagonal matrix and E is a matrix of eigenvectors of A arranged as columns. Thus, we can select P where each row of P is an eigenvector of CC^T. Now, we can rewrite C_Y in terms of P and D.

$$\begin{aligned} C_Y &= \frac{1}{n-1} PAP^T \\ &= \frac{1}{n-1} P(P^T DP)P^T \\ &= \frac{1}{n-1} (PP^T)D(PP^T) \qquad \rightarrow \qquad \textbf{(11)} \\ &= \frac{1}{n-1} (PP^{-1})D(PP^{-1}) \\ &= \frac{1}{n-1} D \end{aligned}$$

It is obvious that P is the matrix that can diagonalize C_Y. The eigenvalues of C_Y (diagonal values in D) are the variances of C, and the row vectors of P corresponding to the largest eigenvalues are the principal components of C.

$$t_N = \sum_{j=1}^{N} d_j \Big/ \sum_{j=1}^{1088} d_j \qquad \rightarrow \qquad \textbf{(12)}$$

where d_j is the jth diagonal value of D, and t_N represents the percentage of the original feature matrix which the selected PCs account for. We choose a cut-off value t_M=0.7. N is the smallest integer, for which $t_N > t_M$. In this, we therefore select the first six principal components from $P(p_1,p_2,p_3,p_4,p_5,p_6)$ as new feature vectors.

Tables:

System	*TP*	*FP*	SN^a	PPV^b
DragonGSF	9	14	64.2	39.1
FirstEF	9	12	64.2	42.9
Eponine	9	16	64.2	36.0
PCA-HPR	9	11	64.2	45.0

Table 1: Performance comparison of four prediction systems for test set 1.[a]Sensitivity (SN): SN=TP/ (TP+FN). $FN=N_{TSS}$-TP, [b]Positive Predictive Value: S_p=TP/(TP+FP).

System	*TP*	*FP*	S_e (%)^a	S_p (%)^b
DragonGSF	269	69	68.4	79.6
FirstEF	331	501	84.2	39.8
Eponine	199	79	50.6	71.6
PCA-HPR	301	65	76.6	82.2

Table 2: Performance comparison of four prediction systems for test set 2.