Promoter Analysis with Wavelets and Support Vector Machines

Makihiko Sato*

Department of System and Information Engineering
Maebashi Institute of Technology, 460-1 Kamisatori-Cho,
Maebashi-City 371-0816, Japan

Abstract

Promoters are key control regions for the transcription regulation of genes, usually lying upstream of the genes they control. Promoter prediction is worthwhile not only for the detection of orphan genes but also for understanding the mechanisms that regulate gene expression. Promoter prediction therefore remains one of the primary challenges subjects in bioinformatics in the post-genome era. Many methods are used for promoter prediction, such as the presence of the CpG islands, sequence motifs of transcription factor binding sites, and the statistical and chemophysical properties in the vicinity of transcription start sites. Among these strategies, we have focused on a method which employs wavelet analysis and support vector machine for promoter prediction. The wavelet analysis is based on localized wave packets characterized by both a range of frequency and a location. In our scheme, information from promoter and non-promoter regions is converted to wavelet space as a positive and a negative set, respectively, and the 2 sets are subsequently used to train a support vector machine. Finally, the support vector machine is utilized for promoter prediction. In this study, we improved the coding method of our prediction strategy and analysed a new set of test data.

"Keywords: promoter prediction; discrete wavelet transformation; support vector machine"

1. Introduction

In the recent decade, a large number of genomes have been sequenced and annotated. Annotation is a crucial component of genomics, and in this area, the development of computational methods for promoter prediction has been a particularly challenging and significant issue [1, 2]. Many methods for promoter prediction have been devised in the post-genome era, and promoter prediction remains important not only for the detection of rarely expressed genes but also for the analysis of the regulatory mechanism of the gene expression [3].

Promoters are key control regions for the transcription regulation of genes, and are usually located upstream of the genes they regulate. Understanding the mechanisms regulating gene expression and identifying the key
regulatory elements is a major challenging in molecular biology [4]. The most common promoter prediction methods are based on sequence similarity or motif finding algorithms [3, 5], their performance of which has been previously evaluated [6]. An alternative approach to promoter prediction is based on the fact that properties of promoter regions are different from those of the other regions of DNA. In general, this method works well in the context of for a particular property of promoter, such as relative DNA stability or C+G rich properties, i.e. CpG islands; where both properties are related to each other. CpG islands are known to preferentially occur at the transcriptional start sites of genes, particularly in the case of housekeeping genes [7, 8].

We previously described a method based on this promoter property that combines discrete wavelet transformation (DWT) and support vector machines (SVMs) [9]. In that study we applied the Daubechies wavelet type basis and, unexpectedly, were unable to detect a difference among different wavelet transformations. In the current study, we applied another type of wavelet basis, B-spline, in addition to the Daubechies type. Moreover, we adopted the Database of Transcriptional Start Sites (DBTSS) instead of the Eukaryotic Promoter Database (EPD).

2. Wavelet analysis

A wavelet is a wave packet of finite location and specific frequency, which rapidly becomes zero at both ends of this location. Wavelet transform is a multi-resolution analysis for signal processing, such as Fourier transform. An essential difference between Fourier and wavelet transform is that Fourier is based on sine and cosine functions that localize in frequency but not in location, whereas wavelet transformation is based on wave packets that localize both frequency and location.

The wavelet transform can be used to analyze a time series that contains highly variable power over a wide range of frequencies [10]. The mathematical introduction to the wavelet theory was shown by Daubechies [11], and the biologists-oriented introduction has been described by Hirakawa et al. [12] and Lio et al. [13]. Typically, wavelets are used for data compression, image compression, noise reduction, radar, earthquake prediction and feature extraction. In recent years, wavelets have been applied to a large variety of biological signals [14, 15], including the detection of patterns in DNA sequences [16, 17].

Mathematically, wavelets are defined as continuous functions, and wavelet transforms are defined by the continuous integration of a signal and the wavelet functions. It is computationally impossible, however, to handle continuous functions and values. Similar to a discrete Fourier transform, a practical wavelet transform is manipulated in a discrete way. In discrete wavelet transform (DWT), a signal is projected onto a wavelet basis function that decomposes the signal into its wavelet coefficients at different scales [18]. Thus, discrete wavelet functions are represented by matrix forms with specific scales and translation values, and integration is replaced by multiplying the wavelet matrix and sampling signals. Typically, the scales used are powers of 2, referred to as a two-scale relation.

In order to investigate the properties of wavelet bases, we have applied 4 orthogonal Daubechies wavelet basis functions (Haar basis (Daubechies 2), Daubechies 4, Daubechies 6, and Daubechies 8) and 3 biorthogonal B-spline wavelet family of orders 103, 202, and 301. The wavelet basis is defined by 2 functions: a scaling function (\(\phi\)) and a wavelet function (\(\psi\)). The relation of the next scale of child functions and the 2 mother functions are expressed as two-scale relation and defined as follows:

\[
\phi^j_i(t) = \sqrt{2^j} \phi(2^j t - i), \quad j = 0,1,\ldots \text{ and } i = 0,1,\ldots,2^j - 1 \\
\psi^j_i(t) = \sqrt{2^j} \psi(2^j t - i), \quad j = 0,1,\ldots \text{ and } i = 0,1,\ldots,2^j - 1
\]

3. Support Vector Machine

Support vector machines (SVMs) are classification and regression methods with supervised learning. SVMs map input vectors nonlinearly to a higher dimensional space where a maximal separating hyper plane is constructed [19, 20]. The separating hyper plane maximizes the distance between 2 different classes of training data. Cover’s theorem insists that a linearly inseparable classification problem can be transformed to be linearly separable by a nonlinear projection onto a sufficiently higher dimensional space [21].

For classification SVMs, each training data is labelled +1 or -1, according to whether the training data is positive or negative, respectively. Separation with a hyper plane stipulates that all of the training data must occur at some
distance from the hyper plane in the mapped space. Once the training is completed, classification is based solely on the positions of the data relative to the separating hyper plane.

4. Methods

Our promoter recognition methods consist of the following steps: test data collection; DWT; SVM training; and evaluation. In the test data collection step, we prepare positive (promoter) and negative (non-promoter) DNA data sets for both training and testing. In the DWT step, these DNA data are transformed into wavelet data after coding to numerical data. The SVM training is performed for positive and negative data after labelling as +1 and -1, respectively. Finally, the other positive and negative data are evaluated using the training SVM.

4.1 Data sets

For the positive data set, the promoter segments of *Homo sapiens* were extracted from the Database of Transcriptional Start Sites (DBTSS) [22]. DBTSS represents the exact positions of transcriptional start sites (TSSs) in the genome based on their unique experimentally validated TSS sequencing method (TSS Seq). Of the 101436 entries of the *H. sapiens* promoter sequence, sequences containing an undefined nucleotide character (‘N’) were omitted, leaving 101420 entries. These sequences were divided into groups of 600. We then randomly selected 2 data sets for the training and test set. For each sequence, we extracted sequences of 512 nt in length (512 = 2^9; 384 nt are upstream of TSS) because of the bi-scale feature of DWT.

Negative data were taken from human chromosome 21 sequences from the DNA Data Bank of Japan (DDBJ) [23]. 600 DNA sequences of 512 nt in length were successively selected from a randomly selected portion of the chromosome for a single negative data set. Three negative data sets were prepared with corresponding positive data sets.

4.2 Discrete Wavelet Transformation (DWT)

Wavelets transformations require numerical sampling data as input. Therefore, the 4 DNA bases - adenine, thymine, guanine, and cytosine – must be converted to a numeric value using a coding rule. We adopted the following coding: [GC] = 1 and [AT] = 0 (or -1), which focuses on the GC content of the promoter, as we previously described [9] and which is in accordance with the work of Lio and Vannucci (2000) who demonstrated wavelet variance decomposition of genomic sequences. Note that this coding does not distinguish between G and C, or between A and T. Four Daubechies type of wavelets, Haar (Daubechies 2), Daubechies 4, Daubechies 6, and Daubechies 8, and 3 Bi-spline wavelet types, Bs-103, Bs-202, and Bs-301, were applied to the test data. For our experiments, we use the wavelet library in the GNU Scientific Library (GSL) [24].

4.3 Support Vector Machine (SVM)

The SVMs were trained using 1 positive data set and 1 negative data set. The segments in the positive data set were labeled as +1, and those in the negative data set as -1. For the SVM, we adopted SVM\textsuperscript{light} - a freely available implementation of SVM in C [25]. We used this SVM with default parameter settings and attempted to apply 4 kernel functions. The 4 kernels are as follows: linear (t0); polynomial (t1): (s a*b + c)^d; radial basis function (t2): exp (-gamma ||a-b||^2); and sigmoid (t3): tanh(s a*b + c). The “i option,” which excludes inconsistent training data and retrain the data sets, was not used in this study. During the training process, t3 did not converge. Moreover, a test run of the training SVM, t2 produced unacceptable results, i.e., every value was the same. As a result, 2 SVM pattern kernels remained - t0 and t1.

5. Results and Discussion

To evaluate the performance of our method on the DBTSS test data, the test data were classified using trained SVMs. In all classification tests, 600 segments of positive and negative sets were used. Figure 1 shows the histogram of the classification results for 4 wavelet types with the linear SVM kernel (t0 option). In these figures,
the horizontal axes represent the output of the SVMs and the vertical axes represent the frequency of each value with an interval of 0.2. Figure 1, (e) and (f) show the same results as Figure 1, (a) ~ (d) from another perspective, which compare the 4 different wavelets basis of the positive and negative data, respectively. The Daubechies 4 data are representative of 4 types of Daubechies basis wavelets, although we tested all wavelet functions. The reason is that no difference can be seen among the histogram results of the 4 cases. Similar results were observed in our previous studies [9].

Fig.1. Distribution histograms of the linear kernel SVMs output: Solid lines show the distribution of positive data and dotted lines show that of negative data; (a) Daubechies 4/linear kernel; (b) B-spline 103/linear kernel; (c) B-spline 202/linear kernel; (d) B-spline 301/linear kernel; (e) all positive data/linear kernel; and (f) all negative data/linear kernel. The horizontal axes represent the classification values of the SVM and the vertical axes represent the frequency with an interval of 0.2.
In addition to the Daubechies wavelet types, we have tested 3 B-spline wavelets - B103, B202 and B301. Figure 2 (a) and (b) show the histogram of Daubechies 4 and B-spline 103 wavelets with the polynomial kernel of the SVM. Table 1 shows the classification performance of 4 wavelet types with the linear kernel SVM. Each wavelet result consists of 2 lines: the upper line represents the positive data and the lower line represents the negative data.

Based on these results, B-spline wavelet types perform slightly better than the Daubechies types. The features of the histograms are similar to each other. In the B-spline results, the B202 appears to be the best, although the difference is small. Compared to our previous work, in which we achieved 91.1% TP and 98.1% TN for the Daubechies 4 wavelet and the linear kernel of SVM respectively, this study has been less successful. We speculate that the decline in performance in the previous study may be due to our use of vertebrate promoters sequence from EPD, while the current study used the H. sapiences promoter sequences from the DBTSS. A second difference between the studies is the TSS location of the TSS in the DNA segments. In the previous study, the TSS was central in the 512 nt segments, whereas in the later study the TSS was located 384 nt from the 5’ end. While TSS locations would ideally have been the same in both studies, the promoter segments of DBTSS sequences ranged from -1000 nt to +200 nt of the TSS. It is possible that this may give rise to a difference between the 2 studies, since the portion downstream of the TSS contains no promoter information.

While we were previously unable to determine why the histograms generated by the Daubechies wavelet types were almost identical to those in our previous study, we can now pose a hypothesis regarding these results.

<table>
<thead>
<tr>
<th>Wavelet</th>
<th>TP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>D4</td>
<td>69.7</td>
<td>30.3</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td>26.7</td>
<td>77.3</td>
</tr>
<tr>
<td>B103</td>
<td>70.5</td>
<td>29.5</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td>23.7</td>
<td>76.3</td>
</tr>
<tr>
<td>B202</td>
<td>71.2</td>
<td>28.8</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td>22</td>
<td>78</td>
</tr>
<tr>
<td>B301</td>
<td>64.5</td>
<td>35.5</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td>20.7</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Fig.2. Distribution histograms of the polynomial kernel SVM output (a) Daubechies 4 (b) B-spline 103

Table 1. Fraction of classification of the linear kernel SVM. Each result shows using 2 lines. The upper line represents the positive data - TP (True Positive) and FN (False Negative). The lower line represents the negative data - FP (False Positive) and TN (True Negative).

<table>
<thead>
<tr>
<th>Wavelet</th>
<th>TP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>D4</td>
<td>79.30%</td>
<td>20.70%</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td>25.30%</td>
<td>74.70%</td>
</tr>
<tr>
<td>B103</td>
<td>79.30%</td>
<td>20.70%</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td>25.00%</td>
<td>75.00%</td>
</tr>
<tr>
<td>B202</td>
<td>79.30%</td>
<td>20.70%</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B301</td>
<td>79.30%</td>
<td>20.70%</td>
</tr>
<tr>
<td>(pos)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(neg)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Fraction of classification of the polynomial kernel SVM. The contents of the table are the same as Table 1.
Daubechies wavelet types are based on the same methodology for their construction, namely, asymmetric bi-scale recursive decomposition. In high frequency regions of each Daubechies wavelet, decomposition must have a different feature. However, it seems that in low frequency regions, the values of decomposition exhibit very similar features and the SVM might extract this similarity during classification. In support of our hypothesis the difference between the histograms of the Daubechies wavelet types and the B-spline wavelet types, although rigorous verification of our hypothesis requires further study.

Subsequent implementations of our scheme will be used to find or classify promoters in long continuous sequences of DNA. In addition, since the wavelets require a scalar value as input, we adopted a rather simple coding in the current iteration. In order to analyze DNA sequence more reliably, more robust wavelet transformations, such as complex wavelet, are desirable.

References