Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides

Jian-Yi Yang[1], Yu Zhou[1], Zu-Guo Yu[1,2]*, Vo Anh[2] and Li-Qian Zhou[1]

[1]School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China.

[2]School of Mathematical Sciences, Queensland University of Technology,

GPO Box 2434, Brisbane, Q 4001, Australia.

Abstract

Background: Promoter region plays an important role in determining where the transcription of a particular gene should be initiated. Computational prediction of eukaryotic Pol II promoter sequences is one of the most significant problems in sequence analysis. Existing promoter prediction methods are still far from being satisfactory.

Results: We attempt to recognize the human Pol II promoter sequences from the non-promoter sequences which are made up of exon and intron sequences. Four methods are used: two kinds of multifractal analysis performed on the numeric sequences obtained from the dinucleotide free energy, Z curve analysis and global descriptor of the promoter/non-promoter primary sequences. A total of 141 parameters are extracted from these methods and categorized into seven groups (methods). They are used to generate certain spaces and then each promoter/non-promoter sequence is represented by a point in the corresponding space. All the 120 possible combinations of the seven methods are tested. Based on Fisher's linear discriminant algorithm, with a relatively smaller number of parameters (96 and 117), we get satisfactory discriminant accuracies. Particularly, in the case of 117 parameters, the accuracies for the training and test sets reach 90.43% and 89.79%, respectively. A comparison with five other existing methods indicates that our methods have a better performance. Using the global descriptor method (36 parameters), 17 of the 18 experimentally verified promoter sequences of human chromosome 22 are correctly identified.

Conclusions: The high accuracies achieved suggest that the methods of this paper are useful for understanding the difficult problem of promoter prediction.

*Corresponding author, e-mail: yuzg1970@yahho.com

Key words: Promoter recognition; transcription start site; multifractal analysis; Z curve; global descriptor.

1. Introduction

Promoter region plays an essential role in determining where the transcription of a particular gene should be initiated. Hence, promoter recognition — the computational task of finding the promoter regions on a DNA sequence, is an important problem [**?**]. The accumulation of a huge amount of genome sequence data in recent years makes the annotation process more and more complicated for higher eukaryotes [**?**]. The RNA polymerase II (Pol II) promoter is a key region that regulates differential transcription of protein coding genes. Computational analysis of Pol II promoters may contribute to improved gene identification and to prediction of the expression context of genes [**?**]. There is a need for prediction techniques that can rapidly and accurately evaluate sequences for the presence of promoter sequences [**?**].

Existing promoter prediction methods are still far from being satisfactory [**?**, **?**, **?**]. The performance of many current eukaryote promoter prediction methods has been unreliable with poor specificity or poor sensitivity [**?**]. Many methods predict promoter sequences based on the regulatory sequence elements (RSEs) in them. But the RSEs are short and not fully conserved in the promoter sequences, which results in a high probability of finding similar sequence elements elsewhere in genomes, outside the promoter regions. That is why most of the promoter prediction methods end up predicting a lot of false positions [**?**]. Fickett and Hatzigeorgiou [**?**] performed an evaluation of the different promoter prediction methods on genome DNA and suggested that it would be worth attempting nonlinear recognition methods, such as neural nets or quadratic discriminant analysis. Following this direction, Gangal and Sharma [**?**] applied time series descriptors and machine learning methods to human Pol II promoter prediction and got a higher accuracy compared with other methods; Kanhere and Bansal [**?**] presented a novel prokaryotic promoter prediction method based on DNA stability showing that the changing in the stability of DNA provides a much better clue than the usual sequence motifs.

In this paper, we attempt to recognize the human Pol II promoter sequences from the non-promoter sequences which contain exon and intron sequences. It should be noted that the aim of the present paper is similar to that of Ref. [**?**], but the non-promoter sequences in Ref. [**?**] are made up of coding sequences (CDSs) and intron sequences, while we use an existing database, the Exon/Intron database, to extract non-promoter sequences. We first convert the promoter/non-

promoter sequences into numeric sequences according to the 10 unified free energy parameters [**?**], which have been used to measure the stability of DNA [**?**]. Then a measure representation is introduced for the numeric sequences. Multifractal analysis of the measure is next performed, which results in the first 5 parameters. Analogous multifractal analysis [**?**] is also used on the numeric sequences to achieve another 4 parameters. The Z curve method, which has been used in recent years with some successes [**?**, **?**], yields 96 parameters for the promoter/non-promoter primary sequences. The protein-chain descriptor method was first proposed by Dubchak *et al.* [**?**] to predict protein folding classes. Here we propose a global descriptor for the promoter/non-promoter sequences, which yields 36 parameters for a global description of the primary sequences. Overall, a total of 141 parameters are extracted from these four different methods and categorized into seven groups (methods). Fisher's linear discriminant algorithm shows that the global descriptor method is the most effective when used separately. Complete enumerations of all the possible combinations of these seven methods (120) are tested to find possibly better results with a relatively smaller number of parameters. Numerical results show that the methods with 96 and 117 parameters can produce satisfactory results. Compared with five other existing tools, the higher sensitivity, specificity, accuracy and correlation coefficient demonstrate that the methods proposed here are useful for understanding the human Pol II promoter prediction problem. 17 of the 18 experimentally verified promoter sequences of human chromosome 22 [**?**] are successfully identified by the global descriptor method (with only 36 parameters).

2. Materials

2.1. Original data

We use two different data sets downloaded from two databases. The first set is the human Pol II promoter sequences from Release 90 of the Eukaryotic Promoter Database (EPD) (www.epd.isb-sib.cn). The EPD is an annotated non-redundant collection of eukaryotic Pol II promoters, experimentally defined by a transcription start site (TSS) [**?**]. The EPD is a useful database when one wants to deal with the Pol II promoter prediction problem and it is broadly tested by different prediction tools [**?**, **?**, **?**, **?**, **?**]. A total of 1871 entries of human Pol II promoter sequences with window size of 499 bp upstream and 100 bp downstream of TSS, which is the same as that used in Ref. [**?**], are obtained from EPD. The sequences containing 'N' are manually filtered out, which results in a total of 1856 sequences. The second set is the non-promoter sequences of the human genome. For this data set, we consider using the Exon/Intron Database (EID), which incorporates information on

the exon/intron structure of eukaryotic genes [**?**] (http://hsc.utoledo.edu/bioinfo/eid/index.html, hs35p1.EID.tar.gz). Firstly, the exon/intron sequences with 'n' and length less than 600 are filtered out. Then, we randomly select 1000 intron sequences from the file hs35p1.intrEID and 500 exon sequences from the file hs35p1.exEID. A fragment of length 600 is then selected randomly from each exon/intron sequence with length larger than 600. As the intron sequences are represented by lower-case letters in the file hs35p1.intrEID, we transform them into upper-case letters to be consistent with the promoter and exon sequences.

2.2. Conversion of the original data

Some studies suggested that various properties, such as stability, bendability and curvature, of the region immediately upstream of the TSS differ from that of downstream region [**?**, **?**, **?**]. The upstream region is less stable, more rigid and more curved than the downstream region. Kanhere and Bansal [**?**] predicted the prokaryotic promoter based on such difference in DNA stability. We convert the original sequences into new numeric sequences according to the free energy of dinucleotides. A sliding window with size of 2nt is used and moved one base pair forward each time. The numeric sequences can be smoothed with a larger window size. For more details on the smoothing method, one can refer to Ref. [**?**]. The free energy values corresponding to the 10 unique dinucleotides are taken from the unified parameters proposed in Ref. [**?**]. They are: AA/TT $= -1.00$ $kcal/mol$, AT/TA $= -0.88$ $kcal/mol$, TA/AT $= -0.58$ $kcal/mol$, CA/GT $= -1.45$ $kcal/mol$, CT/GA $= -1.44$ $kcal/mol$, GT/CA $= -1.28$ $kcal/mol$, GA/CT $= -1.30$ $kcal/mol$, CG/GC $= -2.17$ $kcal/mol$, GC/CG $= -2.24$ $kcal/mol$, GG/CC $= -1.84$ $kcal/mol$. The ten values are added by 2.24 $kcal/mol$ (the negative of the smallest free energy) so that all the values are larger than or equal to zero in order to construct a measure from the time series for the multifractal method in the following analysis. For example, the free energy sequence for one of the promoter sequences with a sliding window of size 2nt is given in Figure 1.

3. Methods

3.1. Multifractal analysis (MFA)

Let T_t, $t = 1, 2, \cdots, N$, be the numeric sequence of a promoter/non-promoter with length N. First, we define

$$F_t = \frac{T_t}{\sum_{j=1}^{N} T_j}, \quad (t = 1, 2, \cdots, N) \tag{1}$$

to be the frequency of T_t. It follows that $\sum_{t=1}^{N} F_t = 1$. We define a measure μ on the interval $[0, 1)$ by

$$\mu(dx) = Y(x)dx, \tag{2}$$

where

$$Y(x) = N \times F_t = \frac{T_t}{\frac{1}{N}\sum_{j=1}^{N} T_j}, \quad x \in [\frac{t-1}{N}, \frac{t}{N}). \tag{3}$$

We denote the interval $[\frac{t-1}{N}, \frac{t}{N})$ by I_t. It is easy to see that $\mu([0, 1)) = 1$ and $\mu(I_t) = F_t$. We call $\mu(x)$ the *measure representation* [**?**, **?**] for the numeric sequence of a promoter/non-promoter.

The most common algorithms of multifractal analysis are the so called *fixed-size box-counting algorithms* [**?**]. In the one-dimensional case, for a given measure μ with support $E \subset \mathbb{R}$, we consider the *partition sum*

$$Z_\varepsilon(q) = \sum_{\mu(B)\neq 0} [\mu(B)]^q, \quad q \in \mathbb{R}, \tag{4}$$

where the sum runs over all different nonempty boxes B of a given side ε in a grid covering of the support E, that is,

$$B = [k\varepsilon, (k + 1)\varepsilon). \tag{5}$$

The *mass exponent* $\tau(q)$ is defined [**?**, **?**] as

$$\tau(q) = \lim_{\varepsilon\to 0} \frac{\ln Z_\varepsilon(q)}{\ln \varepsilon} \tag{6}$$

and the generalized *fractal dimensions* [**?**, **?**] of the measure are defined as

$$D(q) = \frac{\tau(q)}{q - 1}, \quad for \;\; q \neq 1, \tag{7}$$

and

$$D(q) = \lim_{\varepsilon\to 0} \frac{Z_{1,\epsilon}}{\ln \varepsilon}, \quad for \;\; q = 1, \tag{8}$$

where $Z_{1,\varepsilon} = \sum_{\mu(B)\neq 0} \mu(B) \ln \mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of $\ln Z_\varepsilon(q)/(q-1)$ against $\ln \varepsilon$ for $q \neq 1$, and similarly through a linear regression of $Z_{1,\varepsilon}$ against $\ln \varepsilon$ for $q = 1$ [**?**, **?**, **?**]. $D(1)$ is called the *information dimension* and $D(2)$ the *correlation dimension* [**?**, **?**].

The concept of *phase transitions* in multifractal spectra was introduced in the study of logistic maps, Julia sets, and other simple systems. Evidence of a phase transition was found in the multifractal spectrum of diffusion-limited aggregation [**?**]. By following the thermodynamic formulation of multifractal measures, Canessa [**?**] derived an expression for the analogous specific heat as

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \tag{9}$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series.

The singularities of a measure are characterized by the *Lipschitz-Hölder exponent* $\alpha(q)$ [**?**], which is related to $\tau(q)$ by

$$\alpha(q) = \frac{d}{dq}\tau(q). \tag{10}$$

Substitution of Eq. (6) into Eq. (10) yields

$$\alpha(q) = \lim_{\varepsilon \to 0} \frac{\sum_{\mu(B)\neq 0} [\mu(B)]^q \ln \mu(B)}{Z_\varepsilon(q) \ln \varepsilon.}. \tag{11}$$

Again, the exponent $\alpha(q)$ can be estimated through a linear regression of $\{\sum_{\mu(B)\neq 0} [\mu(B)]^q \ln \mu(B)\}/Z_\varepsilon(q)$ against $\ln \varepsilon$. The multifractal spectrum $f(\alpha)$ versus α can be calculated according to a relationship known as *Legendre transformation* [**?**]:

$$f(\alpha) = \min_q \{q\alpha(q) - \tau(q)\}. \tag{12}$$

We first construct a measure for the numeric sequences obtained in Section 2.2 according to Eq. (2), then analyze the measure with the above multifractal method. The $D(q)$, C_q, $\alpha(q)$ and $f(\alpha)$ curves for one of the promoter, exon and intron sequences are shown in Figure 2. We select 5 parameters from MFA to distinguish between promoter and non-promoter sequences: $D(2)$, C_1, C_{max} (the maximum value of C_q), $\Delta\alpha = \alpha_{max} - \alpha_{min}$ and $\Delta f = f(\alpha_{max}) - f(\alpha_{min})$.

3.2. Analogous multifractal analysis ($AMFA$)

Analogous multifractal analysis is similar to *multiaffinity analysis* which is a useful method in many fields. It was recently proposed in [**?**]. We denote a time series as $X(t)$, $t = 1, 2, \cdots, N$.

First, the time series is integrated as

$$y'_q(k) = \sum_{t=1}^{k} (X(t) - X_{ave})^q, \qquad (q \in \mathbb{Z}_+, k = 1, 2, \cdots, N) \tag{13}$$

$$y_q(k) = \sum_{t=1}^{k} |X(t) - X_{ave}|^q, \qquad (q \neq 0, k = 1, 2, \cdots, N) \tag{14}$$

where X_{ave} is the average over the whole time period and $k \in [1, N]$. Then two quantities $M_q(L)$ and $M'_q(L)$ are defined as

$$M'_q(L) = [\langle |y'(j) - y'(j + L)| \rangle_j]^{\frac{1}{q}}, \qquad (q \in \mathbb{Z}_+) \tag{15}$$

$$M_q(L) = [\langle |y(j) - y(j + L)| \rangle_j]^{\frac{1}{q}}, \qquad (q \neq 0) \tag{16}$$

where $\langle \rangle_j$ denotes the average over j, $j = 1, 2, \cdots, N - L$; L typically varies from 1 to N_1 in which the linear fit is good. From the $\ln L$ vs $\ln M_q(L)$ and $\ln L$ vs $\ln M'_q(L)$ planes, one can determine the relations:

$$M'_q(L) \propto L^{h'(q)} \qquad \text{for } q \in \mathbb{Z}_+, \tag{17}$$

$$M_q(L) \propto L^{h(q)} \qquad \text{for } q \neq 0. \tag{18}$$

Linear regressions of $\ln M'_q(L)$ and $\ln M_q(L)$ against $\ln L$ will yield the exponents $h'(q)$ and $h(q)$ respectively.

The exponent $h(q)$ has a nonlinear dependence on q. When $q = 1$, the methods are just those reported in Refs.[**?**, **?**] and these methods are used to study the length sequences from the complete genomes by Yu *et al.* [**?**]. $M'(L)$ may be assessed to determine long-range correlation [**?**]. From Ref. [**?**], the linear fit to get the exponent $h(1)$ is better than that to get the exponent $h'(1)$. Our numerical results show that the exponents $h(q)$ are more robust than the exponents $h'(q)$, so we suggest to use the exponents $h(q)$. We have used $h(q)$ in clustering the structure of large proteins and it turns out to be a useful method [**?**].

Figure 3 gives an example in applying the AMFA to the free energy sequence of a promoter sequence. It shows a good linear relationship between $\ln M(L)$ and $\ln(L)$. For different values of q, we get the exponents $h(q)$ from linear regressions of $\ln M(L)$ against $\ln (L)$ according to Eq. (18). The exponent spectrum $h(q)$ of the promoter sequence is shown in the right panel of Figure 3. We extract four parameters from AMFA: $h(-2)$, $h(-1)$, $h(1)$ and $h(2)$.

3.3. Z curve (ZC)

The concept of the Z curve representation of a DNA sequence was first proposed by Zhang and Zhang [**?**], and was used to distinguish coding and noncoding DNA sequences [**?**, **?**]. A new system based on ZC, Z CURVE 1.0, for finding protein-coding genes in bacterial and archaeal genomes has been proposed [**?**]. Recently, another new self-training system based on the ZC method, ZCURVE_V [**?**], for recognizing protein-coding genes in viral and phage genomes was reported. In this paper, we apply the ZC method in distinguishing promoter and non-promoter sequences. For convenience, we give a brief description of the methods in Refs. [**?**] and [**?**]. The frequencies of bases A, C, G and T occurring in a promoter/non-promoter sequence with bases at positions $1, 4, 7, \cdots; 2, 5, 8, \cdots; 3, 6, 9, \cdots,$ are denoted by $a_1, c_1, g_1, t_1; a_2, c_2, g_2, t_2; a_3, c_3, g_3, t_3,$ respectively. They are in fact the frequencies of bases at the first, second and third codon positions, which can be called *codon-position-dependent* frequencies of mononucleotides. Based on the ZC [**?**], a_i, c_i, g_i, t_i for each i can be used to construct three coordinates, denoted by x_i, y_i and z_i according to the Z transform [**?**]:

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i), \\ y_i = (a_i + c_i) - (g_i + t_i), \\ z_i = (a_i + t_i) - (g_i + c_i), \end{cases} \tag{19}$$

where $x_i, y_i, z_i \in [-1, 1], i = 1, 2, 3$.

We can use the above 9 parameters in the promoter/ non-promoter problem. We can also consider the *codon-position-independent* frequencies of single bases, which results in the following three coordinates:

$$\begin{cases} x = (a + g) - (c + t), \\ y = (a + c) - (g + t), \\ z = (a + t) - (g + c), \end{cases} \tag{20}$$

where $x, y, z \in [-1, 1], a, c, g$ and t are the frequencies for the bases A, C, G and T in a promoter/ non-promoter sequence, respectively.

In addition to the frequencies of codon-position-dependent mononucleotide, we also consider the frequencies of *phase-specific* dinucleotides. We denote the frequencies of the 16 dinucleotides AA, AC, $\cdots$, and TT occurring at the codon positions 1-2 and 2-3 of a promoter or non-promoter sequence by $p_{12}(AA), p_{12}(AC), \cdots, p_{12}(TT); p_{23}(AA), p_{23}(AC), \cdots,$ and $p_{23}(TT)$, respectively. Us-

ing the Z transform [**?**], the following 24 coordinates can be defined:

$$\begin{cases} x_k^X = (p_k(XA) + p_k(XG)) - (p_k(XC) + p_k(XT)), \\ y_k^X = (p_k(XA) + p_k(XC)) - (p_k(XG) + p_k(XT)), \\ z_k^X = (p_k(XA) + p_k(XT)) - (p_k(XG) + p_k(XC)), \end{cases} \tag{21}$$

where $x_k^X, y_k^X, z_k^X \in [-1, 1]$, $p_k(XY) = n_k(XY)/[n_k(XA) + n_k(XC) + n_k(XG) + n_k(XT)]$, $n_k(XY)$ are the occurring times of dinucleotides XY, X, Y=A, C, G, T, $k = 12, 23$.

We can also consider the frequencies of phase-specific dinucleotides and the frequencies of *phase-independent* dinucleotides. For this purpose, a sliding window with size 2nt is used and moved forward one base each time to count the number of times of the occurring dinucleotides. With this method, 12 new coordinates can be defined:

$$\begin{cases} x^X = (p(XA) + p(XG)) - (p(XC) + p(XT)), \\ y^X = (p(XA) + p(XC)) - (p(XG) + p(XT)), \\ z^X = (p(XA) + p(XT)) - (p(XG) + p(XC)), \end{cases} \tag{22}$$

where $x^X, y^X, z^X \in [-1, 1]$, $p(XY) = n(XY)/[n(XA) + n(XC) + n(XG) + n(XT)]$, $n(XY)$ are the occurring times of dinucleotides XY, X, Y=A, C, G, T.

Gao and Zhang [**?**] compared various algorithms for recognizing short coding sequences of human genes and they defined 48 quantities, which were the frequencies of *phase-dependent* tri-nucleotides. In Ref. [**?**], Gao and Zhang used a sliding window with size 3nt and the window was moved forward three bases each time to count the frequencies for the 64 tri-nucleotides. Now we move forward the sliding window with size 3nt one base each time. The definition for the 48 coordinates is

$$\begin{cases} x^{XY} = (p(XYA) + p(XYG)) - (p(XYC) + p(XYT)), \\ y^{XY} = (p(XYA) + p(XYC)) - (p(XYG) + p(XYT)), \\ z^{XY} = (p(XYA) + p(XYT)) - (p(XYG) + p(XYC)), \end{cases} \tag{23}$$

where $x^{XY}, y^{XY}, z^{XY} \in [-1, 1]$, $p(XYZ) = n(XYZ)/[n(XYA) + n(XYC) + n(XYG) + n(XYT)]$, $n(XYZ)$ are the occurring times of trinucleotides XYZ, X, Y, Z=A, C, G, T. The difference between Ref. [**?**] and here is in the calculation of $n(XYZ)$; the present method can be regarded as a *phase-independent* method.

3.4. Global descriptor of promoter/nonpromoter sequence (GD)

Dubchak *et al.* [**?**] proposed a method for predicting protein folding classes based on a global protein chain description. The protein-chain descriptor includes overall composition, transition,

and distribution of amino acid attributes. Similar methods have also been used in Refs. [**?**, **?**, **?**, **?**]. In this paper, we propose the global descriptor of promoter/non-promoter sequences.

The global description contains three parts: composition ($Comp$), transition ($Tran$) and distribution ($Dist$). In order to explain the method, we suppose that a sequence consists of only two kinds of letters (A and B). The composition is used to measure the frequency of occurrence of each kind of letters in the sequences. For example, for the sequence: BABBABABBABBAABABABBAAABBABABA, there are 14 As and 16 Bs, hence the frequencies for A and B are $100.00 \times 14/(14+16) = 46.67$, $100.00 \times 16/(14+16) = 53.33$, respectively. These two numbers represent the first part of the global description, $Comp$. The second part, $Tran$, characterizes the percent frequency with which A is followed by B or B is followed by A. For example, for the above sequence, there are 21 transitions of this type, that is, $(21/29) \times 100.00 = 72.14$. The third part of the global description, $Dist$, measures the chain length within which the first, 25%, 50%, 75% and 100% of certain type of letters is located, respectively. For example, for the above sequence, the first, 25%, 50%, 75% and 100% of Bs are located within the first, 6th, 12th, 20th and 29th nucleotides, respectively. The $Dist$ descriptor for Bs is thus: $1/30 \times 100.00 = 3.33$, $6/30 \times 100.00 = 20.00$, $12/30 \times 100.00 = 40.00$, $20/30 \times 100.00 = 66.67$ and $29/30 \times 100.00 = 96.67$. Likewise, the $Dist$ descriptor for As is 6.67, 23.33, 53.33, 73.33 and 100.00. As a result, the global description for the above sequence is $(Comp; Tran; Dist) = (46.67, 53.33; 72.14; 6.67, 23.33, 53.33, 73.33, 100.00, 3.33, 20.00, 40.00, 66.67, 96.67)$. A more detailed description of global description of sequences is given in Refs. [**?**, **?**, **?**, **?**, **?**].

The global description for the promoter/non-promoter sequences can be computed by a similar procedure. As the sequences consist of four types of nucleotides (A, C, G and T), there are 4 parameters for $Comp$, 6 parameters for $Tran$ and 20 parameters for $Dist$. Overall, a total of 30 parameters are used to give a global description of a promoter/non-promoter sequence.

The Entropy Density Profile (EDP) model is a global statistical description for a DNA sequence, which employs Shannon's artificial linguistic description for a DNA sequence of finite length like an open reading frame (ORF) [**?**]. Zhu *et al.* [**?**] developed a new non-supervised gene prediction algorithm for bacterial and archaeal genomes based on EDP. Here we describe such method briefly. If $p_i (i = 1, 2, 3, 4)$ are the frequencies for the four types of nucleotides of a promoter/non-promoter sequence, then an EDP vector $S = \{s_i\}$ inferred from $\{p_i\}$ is used to represent the sequence with an emphasis on the information content, where i is the index of the four kinds of nucleotides. The EDP s_i is defined as [**?**]

$$s_i = -\frac{1}{H} p_i \log p_i, \ \ i = 1, 2, 3, 4, \tag{24}$$

where $H = -\sum_{i=1}^{4} p_i \log p_i$ is the Shannon entropy.

It was shown that $P = p_1^2 + p_2^2 + p_3^2 + p_4^2$ is a useful statistical quantity for analysis of DNA sequences [**?**, **?**], which was called a nucleotide composition constraint of genomes [**?**]. As a result, we obtain 6 parameters s_1, s_2, s_3, s_4, H and P from EDP.

Overall, combining the above two description systems, we get 36 parameters for the global descriptor of a promoter/non-promoter sequence.

4. Results and discussions

From the four different methods described above, we get a total of 141 parameters. We will test their contributions in the promoter/ non-promoter problem. Then we will try to combine some of them to see whether better results can be achieved.

For comparison of various methods, a benchmark should be set up. We use Fisher's linear discriminant algorithm [**?**, **?**, **?**] to calculate the discriminant accuracies. We divide all promoter and non-promoter sequences into two sets randomly. A set of 90% of promoter/non-promoter sequences is regarded as a training set, and the set of remaining 10% of promoter/non-promoter sequences as a test set.

Fisher's discriminant algorithm is used to find a classifier in the parameter space for a training set. The given training set $H = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$ is partitioned into $n_1 \leq n$ training vectors in a subset H_1 and $n_2 \leq n$ training vectors in a subset H_2, where $n_1 + n_2 = n$ and each $\mathbf{x_i}$ is a κ-dimensional vector, represented by one point in the κ-dimensional parameter space. Then $H = H_1 \cup H_2$. We need to find a parameter vector $\mathbf{w} = (w_1, w_2, \cdots, w_\kappa)^T$ for the κ-dimensional space such that $\{y_i = \mathbf{w}\mathbf{x_i}\}_{i=1}^{n}$ can be classified into two classes in the space of real numbers. If we denote

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x_i} \in H_j} \mathbf{x_i} \quad j = 1, 2, \tag{25}$$

$$\mathbf{S}_j = \sum_{\mathbf{x_i} \in H_j} (\mathbf{x_i} - \mathbf{m_j})(\mathbf{x_i} - \mathbf{m_j})^T, \qquad j = 1, 2, \tag{26}$$

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \tag{27}$$

then the parameter vector $\mathbf{w}$ is estimated as $\mathbf{S}_w^{-1}(\mathbf{m_1} - \mathbf{m_2})$ [**?**]. As a result, Fisher's discriminant rule becomes: "assign $\mathbf{x}$ to H_1 if $\mathbf{Z}(\mathbf{x}) = (\mathbf{m_1} - \mathbf{m_2})^T \mathbf{S}_w^{-1}[\mathbf{x} - \frac{1}{2}(\mathbf{m_1} + \mathbf{m_2})] > 0$ and to H_2 otherwise" [**?**].