Journal of Theoretical Biology 319 (2013) 1–7

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform

Xuan Zhou [a], Zhanchao Li [a], Zong Dai [b], Xiaoyong Zou [b,*]

[a] School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou 510006, PR China
[b] School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, PR China

HIGHLIGHTS

- A novel pseudo-trinucleotide composition based on DWT was proposed.
- The pseudo-trinucleotide composition was used to model SVM for promoter prediction.
- The model achieved good prediction accuracy, far better than the previous results.

ARTICLE INFO

Article history:
Received 18 July 2012
Received in revised form
20 November 2012
Accepted 21 November 2012
Available online 2 December 2012

Keywords:
DNA sequence representation
Support vector machines
Prediction accuracy

ABSTRACT

In this paper, the discrete wavelet transform was introduced into the trinucleotide compositions and a novel DNA sequence representation technique, namely pseudo-trinucleotide compositions was proposed. The pseudo-trinucleotide compositions based on discrete wavelets transform were then employed to model support vector machines (SVM) for the prediction of promoters. The model was evaluated on the genie dataset, and the overall prediction accuracy (*ACC*) by jackknife validation for the classification of promoters, introns and exons was 82.46%, while the *ACC* for the classification of promoters and unpromoters was 82.18%, which was far better than the previous results. The satisfied prediction result revealed that the pseudo-trinucleotide composition based on discrete wavelet transform was an effective representation method for DNA sequence, and plays a very important role in the prediction of DNA function.

1. Introduction

It is well known that the genetic information from DNA to RNA, and then to protein by triplet codons is the central rule of gene expression. The operon model for genes constituted a perfect picture from expression, regulation and replication of genetic information to achievement of complicated gene functions. With the successful completion of Human Genome Project and the rapid development of modern biological science and technology, gene sequence data are emerging at an explosive pace. Analysis of these data to extract the useful information is a hot topic in modern bioinformatics.

The genome includes coding sequences and regulatory sequences. The reliability of gene coding sequence identification has been greatly improved, but the prediction and characterization of regulatory sequences remains challenging problems. Understanding the regulation of gene expression is an important aspect of studying the gene function.

Promoters, which are in the class of regulatory sequences, are the regions of genomic sequences proximal to the transcription start sites (TSS) that are responsible for the initiation of transcription. The research of promoters will be useful in elucidating regulation and expression mechanism of genes (Wasserman and Sandelin, 2004; Werner, 1999). Owing to the availability of vast amounts of genomic data, it is expensive and time-consuming to detect promoters experimentally and manually. Consequently, there is a need for developing prediction techniques that can rapidly and accurately evaluate sequences for the presence of promoters.

In the past years, computational algorithms such as neural networks, genetic algorithms, and linear discriminant functions have been applied for promoter prediction, and there are many promoter prediction tools that have been developed, including the promoter prediction system based on the KL divergence feature selection strategy (Wu et al., 2007), the promoter QSAR models based on pseudo-folding lattice network (LN) and star-graphs (SG) topological indices (Perez-Bello et al., 2009),

* Corresponding author. Tel.: +86 20 84114919; fax: +86 20 84112245.
E-mail addresses: veego_z@hotmail.com (X. Zhou), zhanchao8052@163.com (Z. Li), ceszxy@mail.sysu.edu.cn (X. Zou).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.11.024

Promoter 2.0 (Knudsen, 1999), McPromoter (Ohler et al., 1999), NNPP (Reese, 2001), FirstEF (Davuluri et al., 2001), Dragon promoter finder (Bajic et al., 2003), PromoSer (Halees et al., 2003), EP3 (Abeel et al., 2008), ProSOM (Abeel et al., 2008), N-SCAN (Gross and Brent, 2006), ARTS (Sonnenburg et al., 2006), PSPA (Wang and Hannenhalli, 2006), PromoterExplorer (Xie et al., 2006), BacPP (de Avila et al., 2011), etc. However, many algorithms are limitation to the predictive sensitivity. To improve promoter prediction, it is necessary to introduce other measures to reduce the number of false negative. Therefore, it is the aim of this paper to develop a novel method reducing false prediction rate and improving the prediction accuracy for promoter prediction.

The use of DNA structural information has a lot of potential for assisting with promoter prediction (Burden et al., 2005). Techniques for sequence structural representation and modeling are the key for biological function prediction. Most of the existing representation features of DNA sequence can only represent DNA sequence to a certain extent and neglect the sequence correlation factors. In our previous work, by considering sequence correlation factors, the pseudo-trinucleotide composition based on physicochemical property of nucleotide was proposed for DNA methylation and good results were achieved (Zhou et al., 2011). To expand the application field of this concept, the pseudo-trinucleotide composition was improved for promoter prediction, and a novel pseudo-trinucleotide composition based on discrete wavelet transform (DWT) was proposed. In addition, the machine learning method of SVM was chosen as the modeling method due to its remarkable generalization performance. Consequently, a novel method was proposed to predict promoters with the pseudo-trinucleotide composition based on DWT as the input parameter for SVM.

2. Material and methods

2.1. Dataset

The dataset was collected from the Genie dataset (Reese et al., 2000) which was built by BDGP (Berkeley Drosophila Genome Project). The Genie dataset consists of DNA sequences of promoters, exons and introns, and each kind of sequences have been divided into 5 subsets for 5-fold cross validation previously. In this paper, the previous 100 sequences of each subset were fragmented from -250 bp to $+50$ bp to constitute a dataset which consists of 500 exon sequences, 500 intron sequences and 471 promoter sequences (the number of promoter sequences in Genie dataset was less than 500). The Genie dataset can be downloaded freely from http://www.fruitfly.org/seqtools/datasets/Human/promoter

2.2. Pesudo-trinucleotide composition

To avoid losing many important information hidden in biological sequences, the pseudo amino acid composition (PseAAC) was proposed (Chou, 2001,2005) to replace the simple amino acid composition (AAC) for representing the sample of a protein. For a summary about its recent development and applications, see a comprehensive review (Chou, 2011). Ever since the concept of PseAAC was proposes in 2001, it has rapidly penetrated into almost all the fields of protein attribute prediction. Because it has been widely used, in 2012 a powerful software called PseAAC-Builder (Du et al., 2012) was established for generating various special modes of PseAAC. Stimulated by the concept of PseAAC, the present paper is attempted to propose the pseudo-trinucleotide compositions for predicting the promoter via the approach of wavelets transform.

2.2.1. DNA walk

It is well known that genome includes the coding information of protein synthesis and the regulation information of gene expression. Since DNA sequences consist of 4 nucleotides (A, C, G, T), the biological functions of DNA are decided by the nucleotide distribution in DNA sequences. Therefore, the digitization of DNA nucleotide sequences plays an essential role in the research of DNA biological functions.

DNA walk (Arneodo et al., 1996) is a commonly used method for the digitization of DNA nucleotide sequences. It comes from the concept of random walk in statistics by mapping DNA sequences to time series with some specific rules. There are different dimensional walks such as one-dimensional, two-dimensional DNA walk according to the dimension of mapping spaces.

One-dimensional walk is fairly simple (Arneodo et al., 1996,1998). In the mapping process, if the nucleotide along DNA sequence was the element in the first group, the mapped value make a forward move, otherwise, the mapped value make a backward move. For example, the AT walk (AT–GC) (in AT walk, the nucleotide A and T were elements in the first group, G and C were elements in another group), AC walk (AC–GT) and AG walk (AG–CT) were all usual one-dimensional DNA walk. However, in the one-dimensional walk, two or three nucleotides were considered as the same element, so the difference between nucleotides was neglected. The two-dimensional walk adopted by Guillermo (Abramson et al., 1999) made up the defect by mapping 4 nucleotides into 4 directions of coordinate axes in a two-dimensional plane. Zhang et al. (2001) proposed a three-dimensional Z curve to map DNA sequences. Li (1991) and Li and Kaneko (1992) mapped the DNA sequences to anomalous random walks in three-dimensional space.

In this paper, a two-dimensional DNA walk method was applied and the mapping formula is as follow:

$$X_i = \begin{cases} +1 & i = A \\ 0 & i = G \text{ or } T \\ -1 & i = C \end{cases} \quad (1)$$

$$Y_i = \begin{cases} +1 & i = G \\ 0 & i = A \text{ or } C \\ -1 & i = T \end{cases} \quad (2)$$

where, i is the nucleotide position in DNA sequences, X_i is the orientation of walk curve at the X-axis, Y_i is the orientation of walk curve at the Y-axis. When the nucleotide is (A) or (C), the walk curve moves 1 or -1 at the X-axis. Similarly, when the nucleotide is (G) or (T), the walk curve moves 1 or -1 at the Y-axis. Therefore, the DNA sequences can be converted into digital sequences. Since the DNA sequences in the dataset consisted of 300 nucleotides, the mapping results were two vectors of digital sequences (X-axis and Y-axis) and each vector consisted of 300 discrete digits. Afterwards, each 300 discrete digits of the X-axis and Y-axis were merged into a digital sequence which consisted of 600 discrete digits.

In order to depict the DNA walk process more distinctly, we made a sample as shown in Fig.1. In the figure the promoter sequence "CGCAGCAAAA……" was processed by two-dimensional walk and the results were "$(-1, 0)$ $(-1, 1)$ $(-2, 1)$ $(-1, 1)$ $(-1, 2)$ $(-2, 2)$ $(-1, 2)$ $(0, 2)$ $(1, 2)$ $(2, 2)$……", and so on.

2.2.2. Discrete wavelets transform (DWT)

Since introduced in the early 1980s, wavelets transform (WT) has become a popular signal analysis tool due to their ability to elucidate simultaneously both spectral and temporal information within the signal (Zhou et al., 2003). WT overcomes the shortcoming of Fourier analysis, which is based on functions that are

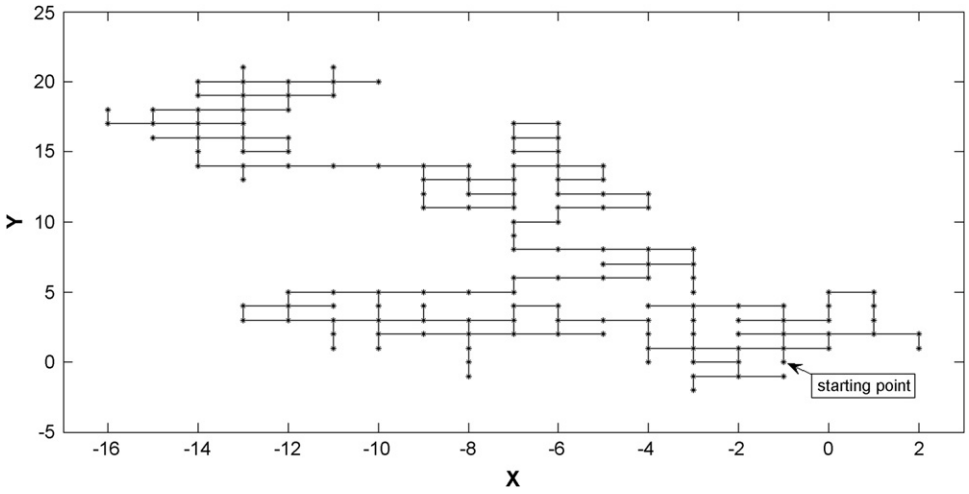

Fig. 1. Graphic expression of DNA walk (take the promoter sequence "CGCAGCAAAATGCACGGGCTTCTGCAGCCCACATGACTTTATTCTGAACGGACACAAGTCTGCTCGCTGGGCCGTTCGCTTTTGGGCCAAAAACACGGCTCCGTCGGTGACTTTTGGCCCGATATTGGCGACCAGAAAACACAAGTGAAAGAGCATTTGGCCAGCCCGGAGAAGCCGAGCTGGGTGGCTTGAGTCTACATGGTTCTCATGTCGCGTTTAAGGCCAGCCCCCTGCACGGTGTGGAGCTTCAATAGCGCAGAGCAGCGTCTACAGCAAAGTACTCCTCCTCACAGACTACCG" as example).

localized in frequency domain, not in time domain, thus leading to location-specific features in the signal being lost (Subramani et al., 2006). A digital signal can be decomposed into many groups of coefficients in different scales with WT, and these coefficient vectors can exhibit characteristics in time domain and frequency domain.

A great deal of research has been carried out on using WT to develop predictors, such as the prediction of protein structural classes (Chen et al., 2012; Li et al., 2009), G-protein-coupled receptor classes (Qiu et al., 2009), enzyme family classes (Qiu et al., 2010), homo-oligomeric proteins (Qiu et al., 2011), membrane protein classes (Rezaei et al., 2008), etc. In all these predictors, the wavelet analysis (continuous wavelet transform (CWT), wavelet power spectrum (WPS), DWT, etc) was used to extract feature of protein sequences based on the physicochemical property of amino acid. In this paper, as a kind of WT which can eliminate the signal noises by discretization of the scale factor and shift factor in WT (Yang, 1999), DWT was adopted to extract feature of DNA sequences based on DNA walk.

By decomposing digit signal into coefficients at different dilations and then removing the noise component from the profiles, DWT analysis can provide local structures of sequences that more effectively reflect the sequence-order effects. The coefficients of DWT can be divided into the approximation coefficient $A^j(n)$, which represents the high-scale and low-frequency components, and the detail coefficient $D^j(n)$, which represents the low-scale and high-frequency components (Qiu et al., 2009). The $A^j(n)$ and the $D^j(n)$ can be expressed as

$$A^j(n) = \sum_{k \in z} h_{k-2n} A^{j-1}(k) \tag{3}$$

$$D^j(n) = \sum_{k \in z} h_{k-2n} D^{j-1}(k) \tag{4}$$

According to the literatures (Chen et al., 2012; Li et al., 2009; Qiu et al., 2009, 2010, 2011; Nanni et al., 2012; Guo et al., 2006; Liu et al., 2005A, 2005B), it is clear that the high-frequency waves contain more noise, while the low-frequency components are functionally more important. Therefore, the use of the low-frequency wavelet coefficients to formulate the sequence can more effectively reflect its overall sequence order effect.

2.2.3. Pesudo-trinucleotide composition based on DWT

Since the amino acid composition can present protein amino acid sequence effectively (Chou, 1995; Chou and Elrod, 2003), the nucleotide composition was proposed to present DNA sequence. DNA sequences only consist of 4 nucleotides (A, C, G, T). If the DNA sequences were represented by single nucleotide composition, the characteristic parameters will be only 4, which are too few as input parameter, so the trinucleotide instead of single nucleotide composition was proposed. However, although the trinucleotide composition includes some order information among the adjacent bases (local structural information), it ignores the sequence order effects of whole DNA sequence (global structural information). In our previous work, the pseudo-trinucleotide composition based on physicochemical property of nucleotide (Zhou et al., 2011) has been proposed to characterize DNA sequences for predicting human DNA methylation. In this work, the pseudo-trinucleotide composition was improved for promoter prediction. The novel pseudo-trinucleotide composition contained not only the sequence order effects of 64-D trinucleotide composition, but also the sequence order effects of whole DNA sequence by processing DNA sequences with DNA walk and DWT.

The WT has a close relation with the DNA sequence characteristic, and it is an effective method for structure analysis of DNA sequence data (Chen et al., 2003). Therefore, DWT was utilized to extract information from DNA sequence. Firstly, the DNA sequences were mapped to digital sequence by DNA walk. Subsequently, the wavelet function and decomposition scale were chosen to perform DWT for the digital sequences, and the low-frequency wavelet coefficients were calculated to obtain new discrete digital sequences that include more comprehensive DNA sequence information. Finally, the new discrete digital sequences were combined with the trinucleotide composition to compose a novel pseudo-trinucleotide composition.

According to equation (6) of Chou (2011), any feature vector of a biological sequence can be formulated as $P=[\psi_1\ \psi_2\ \dots\ \psi_u\ \dots\ \psi_\Omega]^T$, where T is a transpose operator, while the subscript Ω reflects the dimension of the vector and its value as well as the components ψ_1, ψ_2, … will be defined by a series of feature extractions as elaborated below.

The pseudo-trinucleotide composition can be expressed by a vector of $64+\lambda$ dimensions as X. The first 64 components represent the trinucleotide composition, whereas the components

from $64+1$ to $64+\lambda$ reflect the effect of sequence order

$$X = [P_1, \ldots, P_{64}, P_{64+1}, \ldots, P_{64+\lambda}]^{\mathrm{T}} \quad (5)$$

where, P_i is the i-th occurrence frequency of the 64 trinucleotides in the DNA sequence, whereas P_{64+1} to $P_{64+\lambda}$ are the discrete digits by processing DNA sequences with DNA walk and DWT.

The trinucleotide composition contains 64 discrete numbers, each of which reflects the occurrence frequency of one of the 64 trinucleotides in a DNA sequence and includes some order information among the adjacent bases. For the pseudo-trinucleotide composition, with the additional discrete number λ obtained by the DNA walk and DWT process, the sequence order effect of whole DNA sequence can be considered and the characteristic of DNA sequence can be improved. In the modeling of SVM, the 64 trinucleotides as local structural information and the λ additional discrete numbers as global structural information of DNA sequence were complementary for DNA representation to achieve better prediction accuracy.

2.3. Support vector machine (SVM)

In this study, SVM (Vapnik, 1995) was chosen as the modeling method due to its remarkable generalization performance (Liu et al., 2004). The public available LIBSVM software (Chang and Lin, 2001) can be downloaded freely from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

The radial basis function (RBF) was selected as the kernel function. The kernel parameters, including the penalty constant C and the parameters in kernel function (width parameter γ of radial basis function), were optimized by 5-fold cross validation in the modeling process.

Considering that the dataset consists of three subsets: promoter, exon and intron, the prediction is a multi-classification problem. In this study, the one-versus-rest method (Ding and Dubchak, 2001) was adopted to transfer the multi-classification into a series of two-class classifications. For example, for a K-classes problem, there are K binary sub-classifiers needed to be constructed by the one-versus-rest method. The ith sub-classifier is trained by considering all the samples in the ith class as positive samples and all other classes as negative samples. However, the one-versus-rest method may lead to 'False Positive' problem (Ding and Dubchak, 2001). Therefore, the 'winner-takes-all' scheme (Angulo et al., 2003) was utilized, in which the output of each binary classifier was specific numerical values instead of label $+1$ or -1, and the final results of prediction can be given by considering the maximum of the output values.

The structure of prediction model is shown in Fig. 2.

2.4. Evaluation of the predictive performance

The prediction accuracy of each subset (promoter, exon and intron), and the overall prediction accuracy calculated for assessment of the prediction system are given by

$$Accuracy = P_i/N_i \quad (6)$$

$$Overall\ accuracy = \sum_{i=1}^{k} p_i/N \quad (7)$$

where, N is the total number of sequences, k is the class number, N_i is the number of sequences in subset i, p_i is the number of correctly predicted sequences of subset i.

For comparison with other methods, the intron and exon were set as non-promoter for the classification prediction between promoter and non-promoter. The prediction performance was determined by measuring threshold-dependent parameters sensitivity (SE), specificity (SP), accuracy (ACC) and matthews correlation coefficient (MCC) calculated by

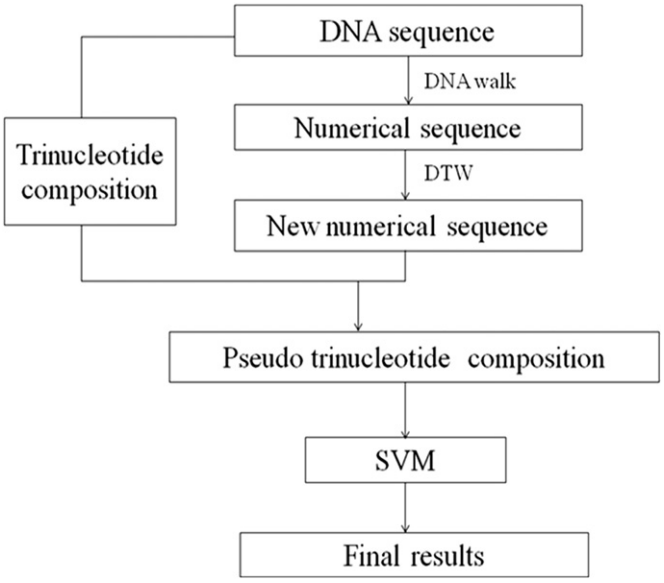

Fig. 2. Flowchart of the current method.

$$SE = TP/(TP+FN) \quad (8)$$

$$SP = TN/(TN+FP) \quad (9)$$

$$ACC = (TP+TN)/(TP+FN+TN+FP) \quad (10)$$

$$MCC = (TP \times TN - FN \times FP)/\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)} \quad (11)$$

where, TP are true positive; FN are false negative; TN are true negative and FP are false positive.

3. Result and discussion

3.1. Optimization of pseudo-trinucleotide composition

In the construction process of pseudo-trinucleotide composition (see Eq. (5)), the λ value is the number of the discrete digits by processing DNA sequences with DNA walk and DWT, decided by the wavelet function and decomposition scale of DWT. The choice of λ value would have a significant impact on the prediction accuracy, therefore, the wavelet function and the decomposition scale of DWT were optimized by 5-fold cross validation.

3.1.1. Optimization of wavelet function

The choice of wavelet function was the crucial factor in wavelet analysis. A variety of wavelet functions were applied to the prediction model and the predicted results were shown in Fig. 3. It can be seen that the haar wavelet function achieved the best prediction accuracy of 0.8103. Therefore, the haar wavelet function was selected as mother wavelet to construct the pseudo-trinucleotide composition.

3.1.2. Optimization of decomposition scale

The decomposition scale j of DWT optimized by 5-fold cross validation was shown in Fig. 4. It can be seen that the prediction accuracy increased gradually with the increase of j. When the value of j was 5, the prediction accuracy reached the maximum of

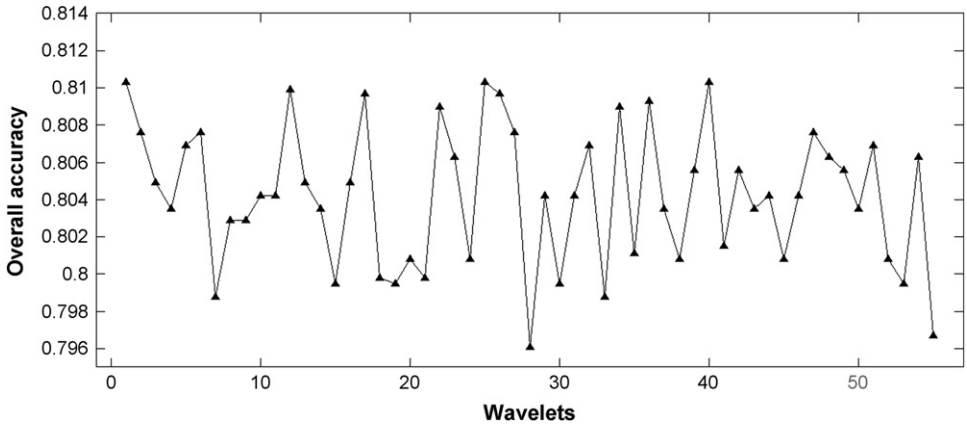

Fig. 3. Overall prediction accuracy based on 55 kinds of wavelet function by 5-fold cross validation. The abscissa 1–55 represents 55 wavelet functions as haar, db1, db2, db3, db4, db5, db6, db7, db8, db9, db10, sym1, sym2, sym3, sym4, sym5, sym6, sym7, sym8, coif1, coif2, coif3, coif4, coif5, bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8, rbio1.1, rbio1.3, rbio1.5, rbio2.2, rbio2.4, rbio2.6, rbio2.8, rbio3.1, rbio3.3, rbio3.5, rbio3.7, rbio3.9, rbio4.4, rbio5.5, rbio6.8 and dmey. The harr wavelet function achieved the best prediction accuracy of 0.8103.

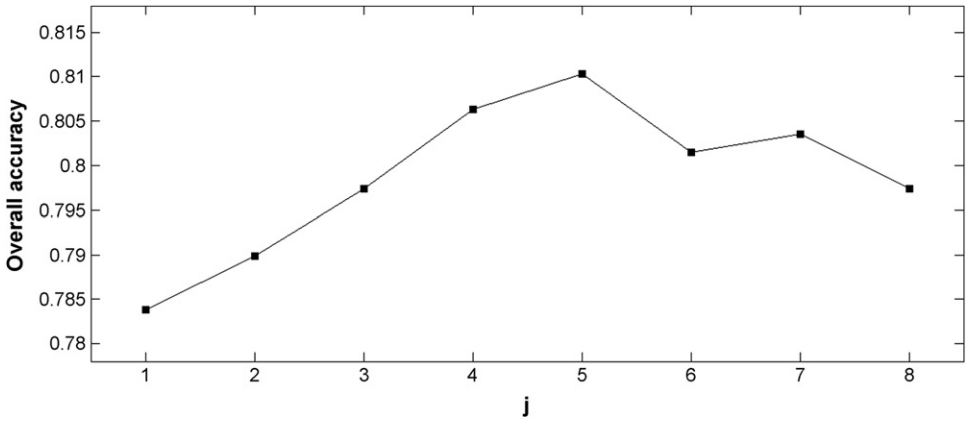

Fig. 4. Overall prediction accuracy based on scale parameter j by 5-fold cross validation. The decomposition scale 5 achieved the best prediction accuracy 0.8103.

0.8103, and then the prediction accuracy decreased. Therefore, the decomposition scale j was optimized as 5.

For the dimension reduction of DWT, the length of data reduces to the half of the original length with every decomposition process. Therefore, when the wavelet function was haar and the decomposition scale was 5, the optimized λ value was 19 (the original 600 discrete digits were obtained by DNA walk).

3.2. Predictive performance

By characterizing the DNA sequence with the optimized pseudo-trinucleotide composition, a model was built for the classification prediction of promoter, intron and exon. In statistical prediction, the following three methods are often used to examine the quality of a predictor: independent dataset test, subsampling test (such as 5-fold or 10-fold cross-validation), and jackknife test. However, as elucidated and demonstrated in (Chou and Shen, 2007, 2008), of the three cross-validation methods, the jackknife test is deemed as the most objective and can always yield a unique result for a given benchmark data set; hence, it has been increasingly used by investigators to examine the accuracy of various predictors (Chou et al., 2011; Fan and Li, 2012; Li et al., 2012; Chou et al., 2012; Zhao et al., 2012; Zia and Khan, 2012). Accordingly, the jackknife test was used in this study to evaluate the predictor. The results were listed in Table 1. The overall prediction accuracy of jackknife test was 82.46%. The prediction accuracy for promoter, intron and exon of jackknife test were 75.58%, 90.40% and 81.00%, respectively.

Table 1
Prediction accuracy on the Genie dataset.

	Accuracy (%)			
	Promoter	Intron	Exon	Overall
Self consistency validation	83.65	87.20	94.80	88.65
5-cross validation	75.16	78.80	88.80	81.03
Jackknife validation	75.58	81.00	90.40	82.46

3.3. Comparison with other methods

To further validate our method, the results of our method were compared with some other methods. In the reported methods, the promoter predictions were mainly about the classification between promoter and non-promoter sequences. Therefore, the half of sequences from intron and exon mentioned above was set as non-promoter for the classification prediction between promoter and non-promoter. The prediction results by jackknife test were listed in Table 2.

From Table 2 it can be seen that the prediction results of our method were much better than other methods. The overall prediction accuracy *ACC* was 82.18%, 11.22% more than that of NNPP method. Furthermore, the reported methods mostly had

Table 2
Comparision with reported methods on the Genie dataset.

Method	*SE*	*SP*	*ACC*	*MCC*
Promoter 2.0 (Knudsen, 1999)	0.4904	0.7800	0.6395	0.2832
McPromoter (Ohler et al., 1999)	0.4798	0.7980	0.6437	0.2938
NNPP (Reese, 2001)	0.5605	0.8500	0.7096	0.4304
FirstEF (Davuluri et al., 2001)	0.3992	0.8320	0.6220	0.2573
Dragon Promoter Finder version 1.5 (Bajic et al., 2003)	0.3800	0.7880	0.5901	0.1844
PromoSer (Halees et al., 2003)	0.4798	0.7780	0.6334	0.2707
Reported pseudo-trinucleotide composition (Zhou et al., 2011)	0.7176	0.8720	0.7971	0.5983
Our method	0.7304	0.9080	0.8218	0.6508

much poorer *SE* than *SP* (the true positive prediction accuracy was relatively poor) which leaded to very poor *MCC*, while the *SE* and *MCC* of our method was 73.04% and 0.6508, far better than those methods.

In addition, the pseudo-trinucleotide composition based on physicochemical property of nucleotide reported in our previous work (Zhou et al., 2011) was also utilized as the feature of SVM to predict promoter for comparison with the novel pseudo-trinucleotide composition based on DNA walk and DWT. The results were also listed in Table 2, and it can be seen that the novel feature improved the prediction accuracy from 79.71% to 82.18%.

To improve the prediction quality of DNA function, it is necessary to characterize the DNA sequence effectively. It is a key problem to incorporate the DNA sequence order effect. The pseudo-trinucleotide composition, a combination of a set of discrete sequence correlation factors and the 64 components of the trinucleotide composition, was proved to have the ability to deal with such a problem. It is anticipated that the pseudo-trinucleotide composition with its mathematical framework and biochemical implication may have a series of impacts on other areas of DNA functions.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors (Chou and Shen, 2007), we shall make efforts in our future work to provide a web-server for the method presented in this paper. The prediction software can be acquired freely on request from the authors.

4. Conclusion

In this paper, the DWT was introduced into the trinucleotide composition and a novel pseudo-trinucleotide composition was proposed to characterize the DNA sequence. The pseudo-trinucleotide composition based on DWT was utilized to model SVM for the prediction of promoters. And the results indicated that the proposed method had the ability to achieve good prediction accuracy. It can be anticipated that the novel DNA sequence representation method may hold a high potential to become a useful tool for predicting other DNA functions.

Acknowledgment

We gratefully acknowledge to the financial support by the National Natural Science Foundation of China (Nos. 20975117, 81171666, 21175158), the Natural Science Foundation of Guangdong Province (10151027501000070), the Scientific Technology Project of Guangdong Province (Nos. 2010A040302001, 2011-B031800257), the Natural Science Foundation of Guangdong Province (10151027501000070), the Ph.D. Programs Foundation of the Ministry of Education of China (No. 20110171110014) and the Zhujiang Scientific Technology Star Project of Guangzhou City.

References

Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., de Peer, Y.V., 2008. Generic eukaryotic core promoter prediction using structural features of DNA. Genome Res. 18, 310–323.
Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., Peer, Y.V. d., 2008. ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. Bioinformatics 24, i24–i31.
Arneodo, A., d'Aubenton, C., Bacry, E., Graves, P.V., Muzy, J.E., Thermes, C., 1996. Wavelet based fractal analysis of DNA sequences. Physica D 96, 291–320.
Arneodo, A., d'Aubenton, C., Bacry, E., Graves, P.V., Muzy, J.F., Thermes, C., 1998. Nucleotide composition effects on the long-range correlation in human genes. Eur. Phys. J. B 1, 259–263.
Abramson, G., Cerdeira, H.A., Bruschi, C., 1999. Fractal properties of DNA walks. BioSytems 49, 63–70.
Angulo, C., Parra, X., Catal_a, A., 2003. A support vector machine for multi-class classification. Neurocomputing 55, 57–77.
Bajic, V.B., Seah, S.H., Chong, A., Brusic, V., 2003. Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. J. Mol. Graphics Modelling 21, 323–332.
Burden, S., Lin, Y.X., Zhang, R., 2005. Improving promoter prediction for the NPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. Bioinformatics 21, 601–607.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255.
Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Chen, C., Shen, Z.B., Zou, X.Y., 2012. Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition. Protein Pept. Lett. 19, 422–429.
Chou, K.C., 1995. A novel approach to predicting protein structural classes in a (20-1)- amino acid composition space. Proteins: Struct. Funct. Genet. 21, 319–344.
Chou, K.C., Elrod, D.W., 2003. Prediction of enzyme family classes. J. Proteome Res. 2, 183–190.
Chen, X.Y., Bao, L.J., Mo, J.Y., 2003. Characterizing long-range correlation properties in nucleotide sequences. Chin. Chem. Lett. 14, 503–504.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Chang, C.C., Lin, C.J. 2001, LIBSVM: A Library for Support Vector Machines, Software Available ftom: 〈http://www.csie.ntu.edu.tw/*cjlin/libsvm〉.
Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nat. Protocol 3, 153–162.
Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.
Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.
Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Davuluri, R.V., Grosse, I., Zhang, M.Q., 2001. Computational identification of promoters and first exons in the human genome. Nat. Genet. 29, 412–417.
de Avila, E.S.S., Echeverrigaray, S., Gerhardt, G.J., 2011. BacPP: bacterial promoter prediction—a tool for accurate sigma-factor specific assignment in enterobacteria. J. Theor. Biol. 287, 92–99.
Du, P., Wang, X., Xu, C., Gao, Y., 2012. PseAAC-Builder: a cross-platform standalone program for generating various special Chou's pseudo-amino acid compositions. Anal. Biochem. 425, 117–119.

Ding, C.H.Q., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.
Fan, G.L., Li, Q.Z., 2012. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids 43, 545–555.
Gross, S.S., Brent, M.R., 2006. Using multiple alignments to improve gene prediction. J. Comput. Biol. 13, 379–393.
Guo, Y.Z., Li, M., Lu, M., Wen, Z., Wang, K., Li, G., Wu, J., 2006. Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30, 397–402.
Halees, A.S., Leyfer, D., Weng, Z.P., 2003. PromoSer: a large-scale mammalian promoter and transcription start site identification service. Nucleic Acids Res. 31, 3554–3559.
Knudsen, S., 1999. Promoter 2.0: for the recognition of Pol II promoter sequences. Bioinformatics 15, 356–361.
Li, W., 1991. Expansion-modification systems: a model for spatial 1/f spectra. Phys. Rev. A 43, 5240–5260.
Li, W., Kaneko, K., 1992. Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. Biophys. Lett. 17, 655–660.
Li, Z.C., Zhou, X.B., Dai, Z., Zou, X.Y., 2009. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415–425.
Liu, H., Wang, M., Chou, K.C., 2005A. Low-frequency Fourier spectrum for predicting membrane protein types. Biochem. Biophys. Res. Commun. 336, 737–739.
Liu, H., Yang, J., Wang, M., Xue, L., Chou, K.C., 2005B. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J. 24, 385–389.
Liu, H.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D., Fan, B.T., 2004. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J. Chem. Inf. Comput. Sci. 44, 161–167.
Li, L.Q., Zhang, Y., Zou, L.Y., Zhou, Y., Zheng, X.Q., 2012. Prediction of protein subcellular multi-localization based on the general form of Chou's pseudo amino acid composition. Protein Pept. Lett. 19, 375–387.
Nanni, L., Brahnam, S., Lumini, A., 2012. Wavelet images and Chou's pseudo amino acid composition for protein classification. Amino Acids 43, 657–665.
Ohler, U., Harbeck, S., Niemann, H., 1999. Interpolated Markov chains for eukaryotic promoter recognition. Bioinformatics 15, 362–369.
Perez-Bello, A., Munteanu, C.R., Ubeira, F.M., De Magalhaes, A.L., Uriarte, E., 2009. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J. Theor. Biol. 256, 458–466.
Qiu, J.D., Huang, J.H., Liang, R.P., Lu, X.Q., 2009. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal. Biochem. 390, 68–73.
Qiu, J.D., Huang, J.H., Shi, S.P., Liang, R.P., 2010. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approachwith support vector machine based on discrete wavelet transform. Protein Pept. Lett. 17, 715–722.
Qiu, J.D., Suo, S.B., Sun, X.Y., Shi, S.P., Liang, R.P., 2011. OligoPred: a webserver for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition. J. Mol. Graphics Modelling 30, 129–134.
Reese, M.G., 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput. Chem. 26, 51–56.
Reese, M.G., Kulp, D., Tammana, H., 2000. Genie—gene finding in *Drosophila melanogaster*. Genome Res. 10, 529–538.
Rezaei, M.A., Abdolmaleki, P., Karami, Z., Asadabadi, E.B., Sherafat, M.A., Abrishami-Moghaddam, H., Fadaie, M., Forouzanfar, M., 2008. Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. J. Theor. Biol. 254, 817–820.
Sonnenburg, S., Zien, A., Ratsch, G., 2006. ARTS: accurate recognition of transcription starts in human. Bioinformatics 22, e472–e480.
Subramani, P., Sahu, R., Verma, S., 2006. Feature selection using Haar wavelet power spectrum. BMC Bioinformatics 7, 432.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for identification of regulatory elements. Nat. Rev. Genet. 5, 276–287.
Werner, T., 1999. Models for prediction and recognition of eukaryotic promoters. Mamm. Genome 10, 168–175.
Wu, S.H., Xie, X.D., Liew, A.W., Yan, H., 2007. Eukaryotic promoter prediction based on relative entropy and positional information. Phys. Rev. E 75, 041908.
Wang, J., Hannenhalli, S., 2006. Amammalian promoter model links cis elements to genetic networks. Biochem. Biophys. Res. Commun. 347, 166–177.
Xie, X.D., Wu, S.H., Lam, K.M., Yan, H., 2006. PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. Bioinformatics 22, 2722–2728.
Yang, F.S., 1999. The Engineering Analysis and Application of Wavelet Transform. Science Press, Beijing.
Zhou, X., Li, Z.C., Dai, Z., Zou, X.Y., 2011. Predicting methylation status of human DNA sequences by pseudo-trinucleotide composition. Talanta 85, 1143–1147.
Zhang, C.T., Wang, J., Zhang, R., 2001. A novel method to calculate the G+C content of genomic DNA sequences. J. Biomol. Struct. Dyn. 29, 333–341.
Zhou, X., Wang, X., Dougherty, E.R., 2003. Binarization of microarray data based on a mixture model. Mol. Cancer Ther. 2, 679–684.
Zhao, X.W., Ma, Z.Q., Yin, M.H., 2012. Predicting protein–protein interactions by combing various sequence—derived features into the general form of Chou's Pseudo amino acid composition. Protein Pept. Lett. 19, 492–500.
Zia Ur, R., Khan, A., 2012. Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix. Protein Pept. Lett. 19, 890–903.