BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 20 2005, pages 3818–3823
doi:10.1093/bioinformatics/bti639

Sequence analysis

Identification of coding and non-coding sequences using local Hölder exponent formalism

Onkar C. Kulkarni, R. Vigneshwar, Valadi K. Jayaraman and Bhaskar D. Kulkarni*

National Chemical Laboratory, Pune 411008, India

Received on March 30, 2005; revised on and accepted on August 17, 2005
Advance Access publication August 23, 2005

ABSTRACT

Motivation: Accurate prediction of genes in genomes has always been a challenging task for bioinformaticians and computational biologists. The discovery of existence of distinct scaling relations in coding and non-coding sequences has led to new perspectives in the understanding of the DNA sequences. This has motivated us to exploit the differences in the local singularity distributions for characterization and classification of coding and non-coding sequences.

Results: The local singularity density distribution in the coding and non-coding sequences of four genomes was first estimated using the wavelet transform modulus maxima methodology. Support vector machines classifier was then trained with the extracted features. The trained classifier is able to provide an average test accuracy of 97.7%. The local singularity features in a DNA sequence can be exploited for successful identification of coding and non-coding sequences.

Contact: Available on request from bd.kulkarni@ncl.res.in

1 INTRODUCTION

With the explosive accumulation of genome sequences, it has become the task of bioinformaticians to annotate a large amount of sequences with a very high degree of accuracy. Annotation includes identification of genes in the genome, assigning putative functions to them and characterizing their boundaries. The algorithms for identification of genes make use of one or more of the several available coding measures. These coding measures incorporate a unique feature/character of the coding sequence, based on which accurate identification of the sequence can be done. Notable among the coding measures, which have been previously exploited are, the codon usage value (Staden and McLachlan, 1982), the Hexamer frequency (Claverie *et al*., 1990) and the mono- and diamino acid usage values (McCaldon and Argos, 1988). Hydrophobicity is another critical parameter for the protein function and has been employed by Tramontano and Macchiato (1986) to take the coding decisions. The base compositional bias in the coding sequences has also been exploited as a coding measure of the sequence (Shepherd, 1981). Silverman and Linsker (1986) use the periodic patterns in the coding sequences revealed by Fourier transform for distinguishing them from non-coding sequences. There are several other global patterns such as dinucleotide frequency, word measure, run measure that have been used as coding measures. Fickett and Tung (1992) have carried out a detailed analysis of various coding measures.

Recently, many algorithms have been developed that use the previously mentioned coding measures for identification of coding and non-coding sequences (Uberbacher *et al*., 1996; Pedersen and Nielsen, 1997; Zhang, 1997; Burge and Karlin, 1997). Zhang and Wang (2000, 2001) have used Z curve representation of DNA sequences to identify coding sequences in *Vibrio cholerae* and yeast genomes by applying the Fischer discriminant algorithm. The Z curve representation of DNA sequence is a 3D space curve, representing the asymmetry in the codon positions with respect to purine/pyrimidine nature of nucleotides, amino/keto nature of nucleotides and strong/weak hydrogen bonding property of nucleotides. Thus the Z curve provides a content measure of the sequence on the basis of which a classification can be done. Methods based on Markov chains have also received wide attention in DNA sequence analysis. Different variations and improvements over conventional Markov models have been implemented in gene classification algorithms. (Salzberg *et al*., 1998; Delcher *et al*., 1999; Borodovsky and McInich, 1993; Lukashin and Borodovsky, 1998). More recently, principles based on non-linear dynamics have been used for the analysis of biological sequences only to reveal interesting statistical behaviors in the sequences. In particular, multifractal analysis has been employed to characterize spatial heterogeneity of the fractal patterns in DNA. A multifractal analysis based on the chaos game representation of DNA sequences (Gutierrez *et al*., 2001) and protein sequences (Yu *et al*., 2004) from complete genomes has been performed. Based on the measure representation of DNA sequences (Yu *et al*., 2001) and the techniques of multifractal analysis, Anh *et al*. (2002) have discussed the problem of recognition of an organism from fragments of its complete genome. Yu *et al*. (2003) proposed the measure representation of linked protein sequence from a complete genome and performed its multifractal analysis. Zhou *et al*. (2005) have used the global features obtained from multifractal analysis of nucleotide sequences to distinguish coding and non-coding sequences. The principles based on non-linear dynamics have also been exploited to detect the presence of long-range correlations in the DNA sequence. Peng *et al*. (1992) use a DNA walk model to discover the presence of long-range correlation in non-coding sequences. Chatzidimitriou-Dreismann and Larhammar (1993) and Prabhu and Claverie (1992) further proved that such correlation also exists in coding sequences. Arneodo *et al*. (1998) and Audit *et al*. (2001) have recently shown the presence of long-range power law correlations in eukaryotic coding sequences.

*To whom correspondence should be addressed.

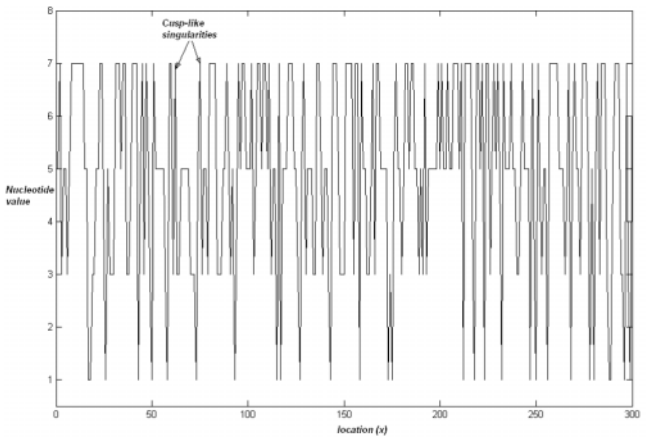

Fig. 1. A time series representation of *B. burgdorferi* DNA coding sequence.

The coding potential of a DNA sequence (the potential of a DNA sequence to encode for a protein) is attributed to the local structures present in the codons. Therefore, the local properties of the DNA sequence will prove to be more informative than the global properties in distinguishing coding and non-coding sequences. Researchers have recently extracted local features from time series in the form of local Hölder exponents to detect outliers in time series (Struzik and Siebes, 2002), to analyze human gait data (Scafetta *et al*., 2003) and to study the influence of progressive central hypovolemia on cardiac interbeat intervals (West *et al*., 2004). We propose to employ information about the local Hölder exponents that capture the local patterns in the time series (numerically encoded DNA sequence) for distinguishing coding and non-coding sequences using binary Support Vector Classification algorithm.

2 METHODS

2.1 Local singularity analysis

Most of the real-life time series data contain step or cusp-like singularities in them (Fig. 1). By singularity, we mean, the rapid changes in the variable values for a very small change in time/position. At those points where singularities are present, the expansion of the time series will contain some components with non-integer powers of time (or position).

Thus, an additional term having non-integer power needs to be included (in Taylor series expansion of the time series) to describe these singularities. The time series around the singularity point x_0 can then represented as (Muzy *et al*., 1994)

$$f(x) = a_0 + a_1(x-x_0) + \cdots + a_n(x-x_0)^n + C|x-x_0|^{h(x_0)}, \qquad (1)$$

where, $a_1, a_2, \ldots, a_n$ denote the expansion coefficients. The exponent $h(x_0)$ is called as the local Hölder exponent and is defined for a point at x_0, as the greatest value of h so that there exists a constant C and an n-th order polynomial, that satisfies the condition (Muzy *et al*., 1994)

$$|f(x) - P_n(x-x_0)| \leq C|x-x_0|^{h(x_0)} \qquad (2)$$

for all values of x in the neighborhood of x_0. It is emphasized that $n < h(x_0) < n+1$ so that $f(x)$ is n times differentiable and its nth derivative is singular in x_0. The local Hölder exponent is a measure of strength of the singularity as well as the regularity of the time series at x_0. The lower value of Hölder exponent at a particular point will reflect a stronger singularity at that point. The estimation of the value of Hölder exponent requires the DNA sequence to be first represented in the form of a numerical time series. Different authors have employed different types of representations in their analysis of DNA sequences (Peng *et al*., 1992; Yu *et al*., 2001, 2004). In this work, we adopt the numerical representation used by Zhou *et al*. (2005), wherein the nucleotide C is represented by the point (1, 1) in 2D space corresponding to its pyrimidine and strong bonding properties; G is represented by (−1, 1) corresponding to it purine and strong bonding properties; A is represented by the point (−1,−1) corresponding to its purine and weak bonding properties; and T is represented by (1,−1) corresponding to its pyrimidine and weak bonding properties. The vectors connecting the origin to the four points (1, 1), (−1, 1), (−1,−1) and (1,−1) will then have rotational angles $\pi/4$, $3\pi/4$, $5\pi/4$ and $7\pi/4$, respectively with the x-axis. The map is then defined as C→1; G→3; A→5; T→7 (Map 1357). For the ambiguous nucleotides in the sequence, we randomly substituted any one of the nucleotides corresponding to the ambiguous representation, with equal probability. With this numerical representation of a DNA sequence, for example, occurrence of C prior to or after the occurrence of T will be more singular and will reflect a lower value of Hölder exponent as compared with the occurrence of C prior to or after the occurrence of G. The change is regular at those locations where the same nucleotide symbols occur together (e.g. AA, TT, GG and CC). The task of detection of the singularities in the DNA series can be conveniently carried out employing wavelet transform (WT). WT has long been known as a vital tool for time series analysis, localized in both time and frequency domains (Strang and Nguyen, 1996). The wavelet transform of $f(x)$ is defined by

$$W_{s,x_0}(f) = \frac{1}{s}\int_{-\infty}^{+\infty} \psi\left(\frac{x-x_0}{s}\right) f(x)\,\mathrm{d}x, \qquad (3)$$

where $\psi(x)$ is a function orthogonal to the polynomial $f(x)$ up to order n, called as the 'wavelet function' or simply 'wavelet'. The scale s fixes the width of wavelet thus adjusting its resolution over the time series. The ability of wavelet transform to reveal even the weaker singularities within the time series by adjusting the scale parameter makes it an indispensable tool for singularity analysis. Applying the wavelet transform to Equation (1), the wavelet coefficient at scale $s \to 0$ for a singularity x_0 can be given as (Struzik, 2000)

$$W_{s,x_0}(f) = Cs^{h(x_0)} \int_{-\infty}^{+\infty} |sx|^{h(x_0)} \psi(x)\,\mathrm{d}x. \qquad (4)$$

If we use a wavelet that has the number of vanishing moments greater than or equal to that of the degree of polynomial $f(x)$, it will filter out the polynomial trends and focus only on the singularities in the time series. A power law proportionality can then be established between the Hölder exponent of a singularity and wavelet transform at that point (Muzy *et al*., 1994; Struzik, 2000),

$$W_{s,x_0}(f) \sim s^{h(x_0)} \quad \text{for } s \to 0^+. \qquad (5)$$

In our work we have used 'Mexican hat' wavelet that satisfies all the required criteria for the purpose on hand. Also, this wavelet is known to converge exponentially to zero at large values of x and thus, applies to a window, which is much larger in size than the scale chosen (Scafetta *et al*., 2003). The Mexican hat wavelet is denoted by

$$\psi(x) = \left(1-x^2\right)\exp\left(\frac{-x^2}{2}\right). \qquad (6)$$

Although Equation (5) can now be used to estimate the Hölder exponents, it is redundant in nature and involves prohibitively large computations. An alternative method proposed by Mallat and Hwang (1992) and Mallat (1999) requires to follow the exponents of scaling along a 'maxima line' to estimate the Hölder exponent of the singularity. The maxima line of a singularity is the line joining modulus maxima of its wavelet coefficients at different scales. A landscape (plot of $\log(s)$ v/s x) of these maxima lines for singularities at all positions gives rises to the wavelet transform modulus maxima

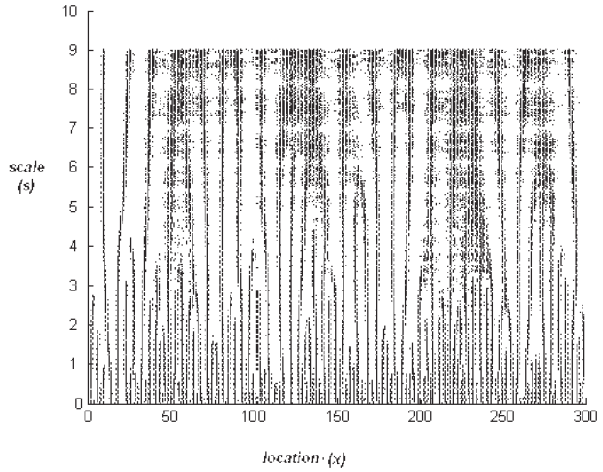

Fig. 2. A WTMM tree for a single coding sequence of *B. burgdorferi*.

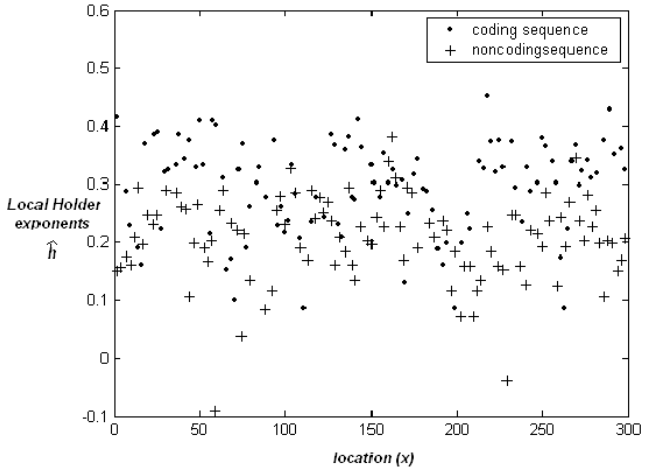

Fig. 3. A Hölder exponent profile for a single coding and single non-coding sequence of *B. burgdorferi* at $s = 1$.

methodology (WTMM) tree (Fig. 2). A more detailed discussion on WTMM tree is provided in Arneodo *et al*. (1995), Struzik and Siebes (2002) and Scafetta (2003).

This WTMM skeleton leads to the definition of a partition function $Z(s, q)$ of q-th moment based on multifractal formalism (Arneodo *et al*., 1995; Muzy *et al*., 1994),

$$Z(s,q) = \sum_{\Omega(s)} |W_{s,x_0}(f)|^q \propto s^{\tau(q)}, \tag{7}$$

where $\Omega(s)$ is the sum of all maxima over scale s and $\tau(q)$ is the scaling exponent that characterizes the power law behavior of this partition function and captures the global distribution of singularities. The Legendre transform of $\tau(q)$ helps in establishing a relationship between itself and global singularity spectrum D_h,

$$h(q) = \frac{d\tau(q)}{dq} \tag{8}$$

$$D_h = qh(q) - \tau(q), \tag{9}$$

where $h(q)$ is the global distribution of Hölder exponents defined at the moment q. The negative values of q capture the weak exponents, whereas the positive values will capture the stronger exponents. The D_h spectrum provides us with global singularity estimates of the time series. However, in certain cases, as ours, the local variations in the time series may prove to be more informative and the local Hölder exponents provide a means to quantify these variations. But the estimation of local Hölder exponents directly from the WTMM representation poses a problem. As the scale value nears zero the wavelet transform will focus more on weaker singularities, and the maxima lines of singularities will become densely packed. Thus a maxima line of one singularity tends to be corrupted by that of the neighboring singularity inducing errors in the estimation of Hölder exponents. To overcome this problem Struzik (1999) suggests an alternative method for estimating the approximate local exponents, in which the singularities are modeled as if they were created through a multiplicative cascading process. The idea is to evaluate a mean of global Hölder exponents ($\bar{h}$) of all the singularities over the chosen scale range using equation

$$\log[M(s)] = \bar{h} \log(s) + c_1, \tag{10}$$

where function $M(s)$ can be viewed as a mean of modulus maxima at a particular scale s defined in terms of partition function [Equation (7)] Struzik and Siebes (2002) as

$$M(s) = \sqrt{\frac{Z(s,2)}{Z(s,0)}}. \tag{11}$$

The approximate local Hölder exponents for singularity at x_0 and at a scale s can be estimated by employing the multiplicative cascade model (Struzik, 1999) as

$$\hat{h}(x_0, s) = \frac{\log(|W_{s,x_0}(f)|) - (\bar{h} \log(s) + c_1)}{\log(s) - \log(s_N)}, \tag{12}$$

where $\hat{h}(x_0, s)$ denotes the approximate value of local Hölder exponents for the singularity at x_0 and scale s. s_N is the maximum available scale for the time series analyzed which is equal to its length N. i.e. $s_N = N$. A typical local Hölder exponent profile for a single coding and non-coding sequence of *B. burgdorferi* is shown in Fig. 3.

2.2 SVM binary classification

Support vector machine (svm) is rigorously based on Vapnik's statistical learning theory (Vapnik, 1995, 1998) and has been employed in several bioinformatics applications (Brown *et al*., 2000; Pavlidis *et al*., 2001; Chris and Dubchak, 2001; Jaakkola, *et al*., 2000; Hua and Sun, 2001; Zhang *et al*., 2003; Bradford *et al*., 2005; Ward *et al*., 2003). The attractive feature of SVM is its excellent generalization capabilities and ability to converge to a single globally optimal solution. For a linearly separable training data the binary SVM builds an optimal hyperplane separating the two classes in the data (Fig. 4). Such an optimal hyperplane maximizes the distance between itself and the nearest data points of each class. For some datasets, especially the biological ones, which are not linearly separable, SVM first maps the input data into a higher dimensional feature space and then constructs a linear hyperplane in the feature space. To avoid the computational problems arising on account of high dimensionality of the feature space, an equivalent kernel function is defined so that the computations can be performed in the input space itself.

Let $\mathbf{x_i} \in \Re$, $i = 1, 2, \ldots, N$ be input training vectors and $y_i \in \{+1, -1\}$ be their corresponding target class (in our case coding and non-coding sequences belong to positive and negative classes, respectively). Let N be the total number of input vectors. The SVM classification problem can then be formulated in terms of a convex quadratic optimization problem (Burges, 1998) as:

$$\max_{\alpha} \left[\sum_{i=1}^{N} \alpha_i - \left(\frac{1}{2}\right) \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right] \tag{13}$$

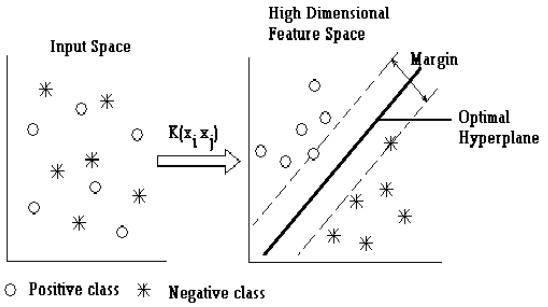

Fig. 4. SVM binary classification.

subject to,

(1) $0 \leqslant \alpha_i \quad i = 1, 2, \ldots, N$
Hard margin problem (No training error allowed)
or
$0 \leqslant \alpha_i \leqslant C_{SV} \quad i = 1, 2, \ldots, N$
Soft margin problem (some training error allowed)
and

(2) $$\sum_{i=1}^{N} \alpha_i y_i = 0,$$

where C_{SV} is a regularization parameter and controls the tradeoff between the SVM complexity and the number of allowable errors in training. $K(\mathbf{x}_i, \mathbf{x}_j)$ denotes the kernel function. We have conducted our simulations with linear kernel and RBF kernel functions defined by Equations (14) and (15), respectively.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \bullet \mathbf{x}_j \tag{14}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\sigma \cdot \frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2}\right) \tag{15}$$

Parameter σ in Equation (15) decides the width of the RBF kernel function.

The target class of any test sequence is determined by using the decision function

$$\Psi(x) = \text{sign}\left(\sum_{i=1}^{nSV} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_j) + b\right). \tag{16}$$

nSV represents the number of training vectors that have a non-zero value of lagrangian multiplier (α_i). Thus, it is evident that knowledge of only a small subset of input vectors is required to form the SVM decision function; therefore, they are called support vectors. The term b in Equation (16) is a bias term that is given by

$$b = -\frac{1}{2} \sum_{nSV} \alpha_i y_i [K(\mathbf{x}_+ \bullet \mathbf{x}_i) + K(\mathbf{x}_- \bullet \mathbf{x}_i)], \tag{17}$$

where $\mathbf{x}_+$ denotes the support vectors of one class and $\mathbf{x}_-$ that from the other.

3 IMPLEMENTATION

3.1 Data

Four genomes from the publicly available NCBI database (ftp://ftp.ncbi.nih.gov/genomes) are used in this work; *Methanobacterium thermoautotrophicum* deltaH (category: Archaebacteria), *Ureaplasma urealyticum* serovar 3 (category: Gram positive Eubacteria low G+C), *B. burgdorferi* B31 (category: Spirochaete), *Buchnera* sp. APS (category: proteobacteria gamma subdivision). In a bacterial genome, the number of non-coding bases (intergenic regions) is less and moreover the intergenic regions might code for structural RNAs. This poses a problem while applying a learning algorithm, such as SVM, as insufficient number of training sequences is available from the non-coding class, which may induce a bias in the results. Therefore, a synthetic method is used to obtain a sufficient number of non-coding sequences for classification (Zhang and Wang, 2001). It is known that the coding potential of a DNA sequence comes from its stringent regular structure in the arrangement of the nucleotides at the three codon positions. (Zhang and Zhang, 1991; Zhang and Chou, 1994). Thus, if this regular structure is disrupted, then the DNA sequence loses its coding potential. For this purpose, coding sequences are first concatenated to form a long coding stretch. Then using a simple randomization algorithm the nucleotides in the coding sequence are shuffled/reordered for sufficient number of iterations. This shuffled sequence can then be labeled as a non-coding sequence. The long stretch of non-coding sequences is cut in lengths equal to that of respective coding sequences to get a non-coding set.

Equation (12) for obtaining the local Hölder exponents is dependent upon the maximum scale available, which is equivalent to the maximum length of a given sequence. This constraint requires the length of all sequences to be nearly constant. Thus to obtain a sequence of uniform length a sliding window approach has been employed. For each sequence, starting from first nucleotide, a window consisting of 300 nt is first selected. This 300 nt long sequence forms our first new sequence. The start and stop positions of the window are shifted by 30 codons, and nucleotides within this new window form the next new sequence. The process of sliding the window is continued till the end of the sequence. This process was repeated for all the coding sequences as well as non-coding sequences. These new sequences of length equal to 300 nt were then used for estimating the local Hölder exponents.

3.2 Estimation of Hölder exponents and classification

A wavelet transform of all the sequences (both coding and non-coding) obtained by the sliding window approach is performed over a range of scales between 1–10 with an interval of 1. Mean global exponent for each sequence is found from the slope of linearly regressed plot of log $M(s)$ versus log(s). The actual estimates of local exponents for each singularity in all sequences were found at a single value of scale, $s = 1$, by using Equation (12). The probability density distribution of these Hölder exponents is then estimated. The features for SVM classification are the probability density values corresponding to the values of Hölder exponents taken at equally spaced increments. For example, the range of Hölder exponents lies between -0.157 to 0.535 in Fig. 5. We divide this range into 40 equally spaced intervals. We now use the probability density values corresponding to the 41 values of the Hölder exponents as features for SVM classification. The training set comprised two-thirds of coding, and two-thirds of the non-coding sequences. The remaining sequences were used as the unseen test samples. A standard 5-fold cross-validation (CV) procedure was employed to estimate the kernel parameter (σ) and the regularization parameter (C_{SV}). The selection of final training parameters was based on average CV error. We use a freely available package, *libsvm*, (Chang and Lin, 2001, http://www.csie.ntu.edutw/~cjlin/libsum) to train the SVM model and for predicting the unseen test sequences. The main steps in the algorithm are listed in Table 1.

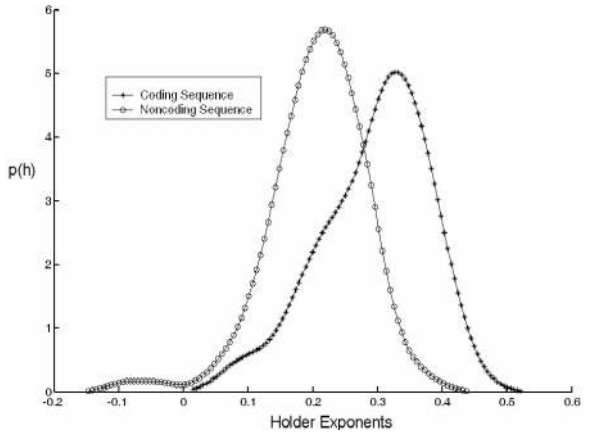

Fig. 5. Probability density plot for a single coding and a single non-coding sequence of *B. burgdorferi*.

Table 1. Main steps in algorithm for classification of coding and non-coding sequences

(1) For each sequence perform the continuous wavelet transform using an appropriate range of scales and a suitable wavelet.
(2) Calculate the partition function at every scale using the summation over all modulus maxima of wavelet coefficients for values of q equal to zero and two (Equation 7).
(3) Calculate the value of M [Equation (11)] for each value of scale and linearly regress the values against the respective scales to obtain the mean local Hölder exponent within the selected scale.
(4) Select a single scale value and estimate the local Hölder exponents at each singularity on the maxima line (Equation 12).
(5) Find the probability density spectrum of these exponents and use these estimates as the features for SVM classification.

4 RESULTS AND DISCUSSION

Current methods of gene recognition employ a variety of biological information for identification of coding regions. In particular, non-uniform codon usage of coding regions is a very well-known and widely used statistical feature. (Fickett, 1982; Galvan *et al*., 2000). Within the coding regions, not all triplets of nucleotides occur with the same probability. The probability of occurrence of a nucleotide is different in each of the three codon positions. Such a preferential codon usage would result in a structured local pattern in the coding regions. In our work the local Hölder exponent density spectrum was employed to capture these local structures effectively. Figure 5 shows the probability density distributions of a single coding sequence and a single non-coding sequence for *B.burgdorferi*. Although in this representative figure for the coding sequence the mean is lower and the standard deviation of the distribution is higher than that of the non-coding sequence, it was found that such a characterization cannot be generalized. If this was so, we could have discriminated the coding/non-coding regions with the information about the mean and standard deviation alone. In general, it was found that such discrimination requires consideration of the entire local Hölder density spectrum along with a state-of-art pattern recognition algorithm, such as SVM. Density spectrum of the extracted local Hölder exponents represents the most informative features of the original DNA series whereas SVM provides the best classification performance when these informative features are supplied to it. Thus we add information in form of 41 probability density values of the local Hölder exponents (covering the entire structural feature of the density spectrum) and employ a machine-learning algorithm, such as SVM to capture and discriminate the trends in the two classes. We implemented two kernels namely linear and RBF kernels. The latter was found to perform better for all the organisms. The classification results are shown in Table 2. To check the effect of using different mappings to convert the DNA sequence into time series, we mapped two organisms with two different mapping namely. C→3; G→5; A→7; T→1 (Map 3571) and C→7; G→1; A→3; T→5 (Map 7135). The results have indicated that there is not much variation in the efficiency with different representations. The ability of Hölder exponents to distinguish the coding sequences from non-coding ones is evident from the average test accuracy of 97.7%. We also have compared our method with Z curve (Zhang and Wang, 2001) and 3-periodic Markov model (Borodovsky and McInich, 1993). Results obtained in this study show that our method gives comparable classification accuracy.

Table 2. Comparison of classification accuracy of Hölder exponent method with Z curve and 3-periodic Markov model

Organism	3-periodic Markov model	Z Curve method	Hölder Exponent method
Borelia burgdorferi	100	97.11	98.20
Ureaplasma urealyticum	99.733	100	97.51
Buchnera sp.	100	99.13	100
Methanobacterium thermoautotrophicum	99.362	98.43	95.36

ACKNOWLEDGEMENTS

We acknowledge the helpful comments of the reviewers that allowed us to improve the content and the presentation of this paper. Financial assistance from Department of Science and Technology, New Delhi, is greatly acknowledged.

Conflict of interest: none declared.

REFERENCES

Anh,V.V. *et al*. (2002) Recognition of an organism from fragments of its complete genome. *Phys. Rev. E*, **66**, 031910.
Arneodo,A. *et al*. (1995) The thermodynamics of fractals revisited with wavelets. *Physica A*, **213**, 232–275.
Arneodo,A. *et al*. (1998) Nucleotide composition effects on the long-range correlations in human genes. *Eur. Phys. J. B*, **1**, 259–263.
Audit,B. *et al*. (2001) Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Lett*., **86**, 2471.
Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.
Borodovsky,M. and McInich,J. (1993) GenMark: parallel gene recognition for both DNA strands. *Computers and Chem*., **17**, 123–133.
Brown,M. *et al*. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci*., **97**, 262–267.
Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Chang,C.C and Lin,C.J. (2001) LIBSVM : a library for support vector machines.

Chatzidimitriou-Dreismann,C.A. and Larhammar,D. (1993) Long-Range correlations in DNA. *Nature*, **361**, 212–213.

Chris,D. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Claverie,J.M. *et al*. (1990) k-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods in Enzymology*, **183**, 237–252.

Delcher,Al *et al*. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., **27**, 4636–4641.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*., **10**, 5303–5318.

Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res*., **20**, 6441–6450.

Galván,P.B. *et al*. (2000) Finding borders between coding and noncoding DNA regions by an entropic segmentation method. *Phy. Rev Lett*., **85**, 1342–1345.

Gutierezz,J.M. *et al*. (2001) Multifractal analysis of DNA sequences using novel chaos-game representation. *Physica A*, **300**, 271–284.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Jaakkola,T. *et al*. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol*., **7**, 95–114.

Lukashin,A. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acid Res*., **26**, 1107–1115.

Mallat,S. and Hwang,W.L. (1992) Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, **38**, 617–643.

Mallat,S.G. (1999) A wavelet tour of signal processing, 2nd Edition. Academic Press, Cambridge.

McCaldon,P. and Argos,P. (1988) Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins*, **4**, 99–122.

Muzy,J.F. *et al*. (1994) The multifractal formalism revisited with wavelets. *International Journal of Bifurcation and Chaos*, **4**, 245–302.

Pavlidis,P. *et al*. (2001) Gene functional classification from heterogeneous data. *Proceedings of RECOMB* .

Pedersen,A.G. and Nielsen,H. (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **5**, 226–233.

Peng,C.K. *et al*. (1992) Long-range correlations in nucleotide sequences. *Nature*, **356**, 158.

Prabhu,V.V. and Claverie,J.M. (1992) Correlations in intronless DNA. *Nature*, **359**, 782.

Salzberg,S.L. *et al*. (1998) Microbial gene identfication using interpolated Markov models. *Nucleic Acids Res*., **26**, 544–548.

Scafetta,N. *et al*. (2003) Hölder exponent spectra for human gait. *Physica A*, **328**, 561–583.

Shepherd,J.C.W. (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. USA*, **78**, 1596–1600.

Silverman,B.D. and Linsker,R. (1986) A measure of DNA Periodicity. *J. Theor. Biol*., **118**, 295–300.

Staden,R. and McLachlan,A.D. (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res*., **10**, 141–156.

Strang,G. and Nguyen,T. (1996) *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesely MA.

Struzik,Z.R. (1998) Removing divergences in the negative moments of the multi-fractal partition function with the wavelet transformation. CWI Report, INS-R9803.

Struzik,Z.R. (2000) Determining local singularity strengths and their spectra with the wavelet transform. *Fractals*, **8**, 163–179.

Struzik,Z.R. and Siebes,A.P.J.M. (2002) Wavelet transform based multifractal formalism in outlier detection and localization for financial time series. *Physica A*, **309**, 388–402.

Tramontano,A. and Macchiato,M.F. (1986) Probability of coding of a DNA sequence: an algorithm to predict translated reading frames from their thermodynamic characteristics. *Nucleic Acids Res*., **14**, 127–135.

Uberbacher,E.C. *et al*. (1996) Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol*, **266**, 259–281.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, NY.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, NY.

Ward,J.J. *et al*. (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 1650–1655.

West,B.J. *et al*. (2004) Influence of progressive central hypovolemia on holder exponent distributions of cardiac interbeat intervals. *Annals of Biomed. Engg*., **32**, 1077–1087.

Witten,I.H. and Frank,E. (2000) Data mining: practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco.

Yu,Z.G. *et al*. (2001) Measure representation and multifractal analysis of complete genome. *Phys. Rev. E*, **64**, 031903.

Yu,Z.G. *et al*. (2003) Multifractal and correlation analysis of protein sequences from complete genome. *Phys. Rev. E*, **68**, 021913.

Yu,Z.G. *et al*. (2004) Chaos game representation and multifractal and correlation analysis of protein sequences from complete genome based on detailed HP model. *J. Theor. Biol*., **226**, 341–348.

Zhang,C.T. and Zhang,R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res*., **19**, 6313–6317.

Zhang,C.T. and Chou,K.-C. (1994) A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences. *J. Mol. Biol*., **238**, 1–8.

Zhang,C.T. and Wang,J. (2000) Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res*., **28**, 2804–2814.

Zhang,C.T. and Wang,J. (2001) Identification of protein-coding genes in the genome of *Vibrio cholerae* with more than 98% accuracy using occurrence frequencies of single nucleotides. *Eur. J. Biochem*., **268**, 4261–4268.

Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis,. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.

Zhang,S.W. *et al*. (2003) Classification of protein quaternary structure with support Vector Machine. *Bioinformatics*, **19**, 2390–2396.

Zhou,L.Q. *et al*. (2005) A fractal method to distinguish coding and noncoding sequences in a complete genome based on a number sequence representation. *J. Theor. Bio*., **232**, 559–567.