GMR
Genetics and Molecular Research
ISSN 16765680

A wavelet-based feature vector model for DNA clustering

J.P. Bao and R.Y. Yuan

Department of Computer Science & Technology, Xi’an Jiaotong University,
Xi’an, China

Corresponding author: J.P. Bao
E-mail: baojp@mail.xjtu.edu.cn

Genet. Mol. Res. 14 (4): 19163-19172 (2015)
Received August 2, 2015
Accepted October 1, 2015
Published December 29, 2015
DOI http://dx.doi.org/10.4238/2015.December.29.26

ABSTRACT. DNA data are important in the bioinformatic domain. To extract useful information from the enormous collection of DNA sequences, DNA clustering is often adopted to efficiently deal with DNA data. The alignment-free method is a very popular way of creating feature vectors from DNA sequences, which are then used to compare DNA similarities. This paper proposes a wavelet-based feature vector (WFV) model, which is also an alignment-free method. From the perspective of signal processing, a DNA sequence is a sequence of digital signals. However, most traditional alignment-free models only extract features in the time domain. The WFV model uses discrete wavelet transform to adaptively yield feature vectors with a fixed dimension based on the features in both the time and frequency domains. The level of wavelet transform is adjusted according to the length of the DNA sequence rather than a fixed manually set value. The WFV model prefers a 32-dimension feature vector, which greatly promotes system performance. We compared the WFV model with the other five alignment-free models, i.e., k-tuple, DMK, TSM, AMI, and CV, on several large-scale DNA datasets on the DNA clustering application by means of the K-means algorithm. The experimental results showed that the WFV model outperformed the other models in terms of both the clustering results and the running time.

Key words: DNA clustering; Discrete wavelet transform; Alignment-free model

INTRODUCTION

Extracting valuable information from a large amount of biological data is the primary goal of bioinformatics. A clustering technique is commonly applied to DNA and protein sequences, by which we can explore the inherent relationships between biological species (Liu et al., 2006). DNA sequences in the same cluster are regarded as homologous. From this, we can determine the function of unknown genes and lay the foundation for further research.

The similarity metric between DNA sequences is one of the keys to DNA clustering. Generally, there are two basic ways of establishing similarity between DNA sequences. One is the alignment-based method, and the other is the alignment-free method. The alignment-based method compares two or more sequences based on string matching methods, which are time consuming. Moreover, it is difficult for an alignment-based method to cluster DNA sequences of varying length. The alignment-free method calculates similarity quickly by converting DNA sequences into unified feature vectors. A good feature vector retains important information and suppresses noise.

Since a DNA sequence can be translated into a sequence of digital signals, the feature vector can be built in time or frequency domains. However, most traditional alignment-free models, such as k-tuple (Vinga and Almeida, 2003), DMK (Wei and Jiang, 2010), TSM (Shi and Huang, 2012), AMI (Bauer et al., 2008), and CV (Qi et al., 2004) models build their feature vectors only in the time domain, i.e., they use direct word sequences.

This paper presents a model to create a DNA feature vector called the wavelet-based feature vector (WVF) model that is based in both the time and frequency domains. The main characteristics of the WFV model are as follows: 1) The WFV model exploits the discrete wavelet transform (DWT) to adaptively decompose and extract features of a DNA sequence according to its length rather than the DNA microarray data. As a result, sequences of different length can be converted into the same-sized feature vector. 2) The WFV model is able to achieve better clustering results with small feature vectors (only 32 dimensions).

Our experiments showed that the WFV model is superior to the five alignment-free models mentioned above.

The paper is organized as follows: The Related Work section describes relevant work on alignment-free and wavelet-based models. The Methods section introduces the procedure of WFV. Experimental proof that WFV surpasses other models is presented in the Results section. Finally, our interpretation of the results is summarized in the Conclusions section.

RELATED WORK

Most of the traditional alignment-free models build feature vectors based on the probability distribution of words, i.e., short consecutive DNA characters. For example, k-tuple (Vinga and Almeida, 2003) adopts the sliding window k to segment DNA sequences. It counts the word frequency that is used to produce the feature vectors with fixed dimension 4^k. However, k-tuple cannot fully describe all the information in DNA. DMK (Wei and Jiang, 2010) brings in the distribution information of DNA sequences based on the position of each word. By analyzing the classifications of nucleotide bases, Shi and Huang (2012) transformed a DNA sequence to three DNA sequences, and counted word frequencies in them to create the feature vectors. Bauer et al. (2008) used the average mutual information to represent the DNA sequence information, called the AMI model. Qi et al. (2004) proposed a method named CV, which applies word frequency and the Markov chain theory to phylogenetic reconstruction. Chang et al. (2014) also used the Markov model and k-word distributions to compare HIV and HEV genome sequences. Heyne et al. (2012) mapped RNA sequence structure information into a graph and clustered the RNAs by means of a graph kernel. Leimeister and Morgenstern (2014) compared DNA sequences based on the longest common substrings with k mismatches.

Eisen (1998), Yi et al. (2007), and Hatfull et al. (2010) clustered genomes by various types of statistical information from DNA microarray data. Bonham-Carter et al. (2014) summarized 14 alignment-free genetic sequence comparison methods. However, all these methods only utilize features from short string and probability information, without any features from the transformed space, such as frequency domain.

Fourier transform is often exploited to extract features in the frequency domain. SRF (Sharma et al., 2004) and SBARS (Pyatkov and Pankratov, 2014) adopt Fourier transform to identify dispersed and tandem DNA repeats. The latter is more suitable for analyzing long sequences and searching for extended homologous fragments. Satsuma (Grabherr, 2010) is a sequence alignment program that finds sequence matches through cross-correlation implemented by fast Fourier transform (FFT).

Wavelet transform is faster and more efficient than Fourier transform in capturing the essence of data (Liò, 2003). There is a growing interest in using wavelet transform to analyze biological sequences and molecular biology-related signals. Machado et al. (2011) studied human DNA in the context of signal processing by encoding nucleotides as complex numbers. Wang et al. (2010) and Du et al. (2006) proposed peak detection algorithms based on the wavelet theory. Pique-Regi et al. (2007) used wavelet footprints to represent the DNA copy number. Liu (2007) described a feature selection method based on wavelet analysis and a genetic algorithm. Abbasi and Rasi (2011) applied discrete wavelets to decrease the output spectrum noise so as to identify exonic regions.

Li et al. (2004) presented a new prediction system with a clustering algorithm, in which the wavelet analysis is performed to identify proteins in a cluster. Alexandrov et al. (2009) presented a biomarker discovery algorithm based on DWT. Nanni et al. (2012) combined different feature reduction approaches to improve classification performance, including tree wavelet. Nanni and Lumini (2011) also compared a set of orthogonal wavelet mother functions to extract the features from the microarray data. Rashid and Maruf (2011) used a wavelet decomposition technique to reduce features from the DNA microarray data.

METHODS

Motivation

There are many challenges to the improvement of the DNA clustering application. For a large amount of data, the alignment-free method is more suitable than the alignment-based method. It is important for the alignment-free model to build the feature vector of a DNA sequence, since a DNA sequence can easily be considered a sequence of digital signals. Therefore, we can analyze the features in the transformed domain, i.e., the time domain and the frequency domain.

Most traditional alignment-free models extract features only in the time domain, which does not adequately describe a DNA sequence. The frequency domain features can be extracted by fourier transform (FT) or wavelet transform (WT). FT addresses a signal accurately in the frequency domain, but it does not have any resolution in the time domain, and its time complexity is $O(N^2)$. WT can address a signal in both the frequency and time domains. FT does not have scale changes, while WT has multi-scale characteristics, which can contact frequency and position. The time complexity of WT can be O(N), where N denotes the size of the data.

Therefore, this paper proposes a WFV model, which is an alignment-free model to build feature vectors in the frequency domain by DWT. The WFV model directly deals with DNA sequences rather than DNA microarray data or mass spectrometry data. To convert sequences of different length into the same size vector, we have also introduced an adaptive method to set the wavelet decomposition level. The WFV model is efficient and suitable for large amounts of data. It is more important that the WFV model achieves better DNA clustering performance than other alignment-free models.

The wavelet-based feature vector model

The input of wavelet transform is a digital signal, but a DNA sequence consists of characters. Therefore, the first step of WFV is to convert a DNA sequence to a digital sequence. A sequence S is defined as a linear succession of N symbols from a finite alphabet. WFV directly converts character sequences to digital sequences according to the following rules:

$$\begin{aligned} CODE[i] &= 0 \textit{ if } S[i]=A, \quad CODE[i] = 1 \textit{ if } S[i]=C \\ CODE[i] &= 2 \textit{ if } S[i]=G, \quad CODE[i] = 3 \textit{ if } S[i]=T \end{aligned} \qquad \text{(Equation 1)}$$

where S is the DNA sequence, CODE is the code corresponding to S, S[i] denotes the i*th* character in S, and CODE[i] denotes the i*th* code in CODE; 1≤i≤N, and N is the length of the DNA sequence. The length of CODE N is equal to the length of the DNA sequence. For example, if the DNA sequence was “ACGTTAGC”, its corresponding code would be “01233021”.

In addition, we have tried other ways to encode DNA sequences, such as word frequency, position distribution information, and classification of nucleotide bases. However, our experimental results show that encoding DNA sequences based on (1) can achieve better clustering results than the other methods.

WFV adopts DWT to obtain feature vectors of DNA sequences. At each level of wavelet decomposition, the signal is decomposed into approximation coefficients (ACs) and detail coefficients (DCs). An AC retains most of the energy of the original signal, which reflects the general characteristics of the original signal. A DC mainly embodies the partial feature of the signal, which denotes the changes of the original signal in detail.

The length of the AC is reduced to half at each level. At the L*th* level of DWT, the length of the AC is reduced to $1/2^L$. As the decomposition level increases, the AC becomes shorter, more information is lost, and the deviation is bigger. Hence, it is crucial to select the proper decomposition level.

WFV adaptively selects the decomposition level according to the length of the DNA sequence. Because each DNA sequence is converted into a feature vector whose length is fixed to M, we can determine the decomposition level using (Equation 2):

$$L = \left\lceil \log_2\left(\frac{N}{M}\right) \right\rceil \qquad \text{(Equation 2)}$$

where N denotes the length of the DNA sequence, M is the fixed length of the feature vector, and L is the decomposition level.

DWT reduces the dimension of CODE. WFV uses the simplest Haar wavelet to create the feature vector of DNA sequence S, i.e.,

$$F(S) = Haar_{AC}(CODE_S, L) \qquad \text{(Equation 3)}$$

where F(S) denotes the feature vector of DNA sequence S, $CODE_S$ is the code of S, and L is the decomposition level. If the length of $CODE_S$ N is less than M×2^L, then WFV puts M×2^L-N zeros at the end of $CODE_S$.

After adaptive L level DWT decomposition, every DNA sequence is converted into a feature vector with fixed length M. The similarity between feature vectors is measured by Euclidean distance, i.e.,

$$E(\mathrm{S}_1, \mathrm{S}_2) = \left\| F(\mathrm{S}_1) - \mathrm{F}(\mathrm{S}_2) \right\| = \sqrt{\sum_{\mathrm{i}=1}^{M} \left| \mathrm{F}_i(\mathrm{S}_1) - \mathrm{F}_i(\mathrm{S}_2) \right|^2} \qquad \text{(Equation 4)}$$

where S_1 and S_2 are the original DNA sequences, and $F(S_1)$ and $F(S_2)$ denote their feature vectors, respectively.

In this paper, we tested the WFV model by the DNA clustering application. Namely, the task was to divide the DNA sequences that belong to the same family into the same cluster as well as possible. Although the data origin was clear, no clustering algorithm could precisely and correctly rearrange all the data into the correct clusters. In particular, for the large-scale collection of DNA datasets, there is much room for improvement with regard to the clustering results.

The pseudo-code

The following lists the pseudo-code for the WFV model.

```
Algorithm 1 The WFV model:
procedure WFV (DNA sequence $S_1$, DNA sequence $S_2$, the length of the feature vector M);
    for all DNA sequence S do;
        CODE←encode(S), by Eq. (1)
```

```
$L \leftarrow level(N,M)$, by Eq. (2)
$CODE,0 \leftarrow CODE$ if $N<Mx2^L$
$F(S) \leftarrow Haar_{AC}(CODE_S,L)$, by Eq. (3)
end for
$E(S_1,S_2) \leftarrow F(S_1), F(S_1)$, by Eq. (4)
return $E(S_1,S_2)$
end procedure
```

We tried to split a DNA sequence into several windows with fixed length and then apply the DWT to the fixed windows, but our tests show that the clustering results from this method were worse than those from the WFV model.

Regardless of the variance in the length of the original DNA sequences, WFV always makes fixed length feature vectors. It is a great advantage to compare DNA similarity by the popular clustering algorithm, such as K-means. Our experimental results show that the WFV model is better than the other models we have tried.

RESULTS

Experiment settings

This paper uses three datasets HOG100, HOG200, and HOG300, which were collected using population-based incremental learning (PBIL). Each dataset was randomly selected from HOGENOM, which contains homologous gene families from microbial organisms. Table 1 lists the details of these datasets. The HOG* dataset contains families that vary from 100 to 300.

Table 1. Details of the datasets.

Dataset	Number of families	Number of DNA sequences in the dataset	Average length of a DNA sequence in the dataset	Dataset size (MB)
HOG100	100	9648	1484	15.1
HOG200	200	22585	1557	37.0
HOG300	300	27825	1448	42.6

We implemented DWT by the PyWavelets module in Python, and the DNA clustering procedure was performed by the K-means algorithm, which is implemented by the SciPy module in Python. As is well known, the K-means algorithm selects initial cluster centers randomly so that the clustering results vary. To even out infrequent results, we executed the K-means procedure 200 times to calculate the average performance.

The clustering results were evaluated by F-measure, which is defined as follows: Let G denote the number of families in the dataset, H denote the number of clusters in the whole clustering result, m denote the number of total sequences in the dataset, m_i denote the number of sequences in family i, c_j denote the number of sequences in the cluster j, and m_{ij} denote the number of sequences that belong to both family i and cluster j.

The precision of cluster j to family i is as follows:

$$p(\mathrm{j},\mathrm{i}) = \frac{m_{ij}}{c_j} \qquad \text{(Equation 5)}$$

The recall of cluster j to family i is as follows:

$$r(\mathrm{j},\mathrm{i}) = \frac{m_{ij}}{m_i} \qquad \text{(Equation 6)}$$

The F-measure of cluster j to family i is as follows:

$$F_1(\mathrm{j},\mathrm{i}) = \frac{2 \times p(\mathrm{j},\mathrm{i}) \times r(\mathrm{j},\mathrm{i})}{p(\mathrm{j},\mathrm{i}) + r(\mathrm{j},\mathrm{i})} \qquad \text{(Equation 7)}$$

The F-measure of the whole clustering result is as follows:

$$F_1 = \sum_{j=1}^{H} \frac{c_j}{m} \max_{i=1}^{G}(F_1(\mathrm{j},\mathrm{i})) \qquad \text{(Equation 8)}$$

Preferred length of feature vectors

The length of the feature vector in the WFV model is a manually set fixed value, which has an important effect on the time complexity and the clustering results. Obviously, the longer the feature vector is, the more computing resources, including central processing unit (CPU) time and memory space, are consumed. We varied the length of feature vectors from 8 to 512 to find the preferred length. Figure 1 illustrates the clustering results against feature vector lengths on the three datasets.

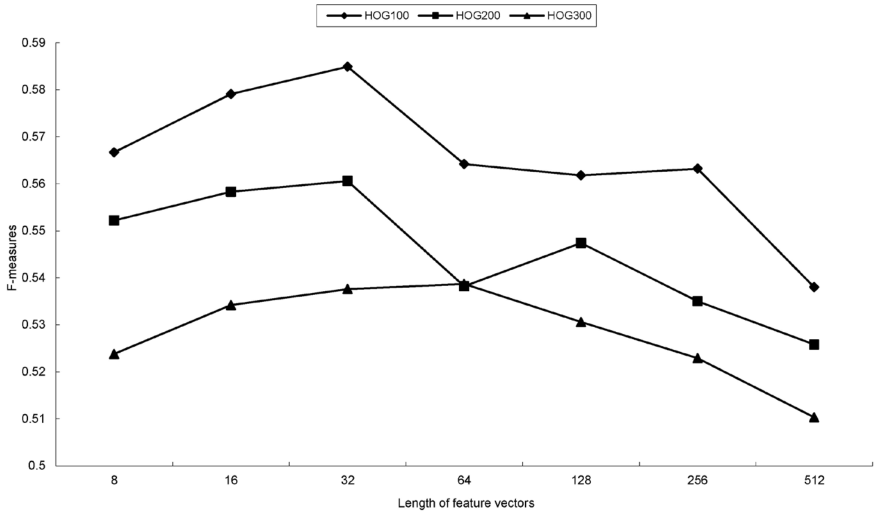

Figure 1. Clustering results of the wavelet-based feature vector (WFV) model against the length of the feature vector on different datasets.

On HOG100 and HOG200, WFV achieved the best clustering result when the length of the feature vector was 32, and the next best when the vector was 16. On HOG300, the WFV achieved the best clustering result when the length of the feature vector was 64, and the next best when the vector was 32. However, the difference of clustering results was very small between 16 and 64. Consequently, 32 was the preferred length of feature vectors in our tests.

As a result, a longer feature vector may not achieve a better clustering result. A shorter feature vector can reduce computation time, which is very helpful in processing large-scale DNA sequences.

Clustering results

We compared the WFV model with the five other alignment-free models, i.e., k-tuple, DMK, TSM, AMI, and CV. The WFV model fixed the length of the feature vector to 32. For the other five models, the feature vectors were significantly affected by the size of the sliding window. The sliding window for TSM was 2, while for the rest it was 3. Since the length of each codon is three in DNA, it might be beneficial to retain the genetic information of DNA.

Figure 2 illustrates the clustering results from all the models in the F-measure on the three datasets. It is clear that the performance of WFV was best. DMK did not perform quite as well as WFV, but was much better than other modules.

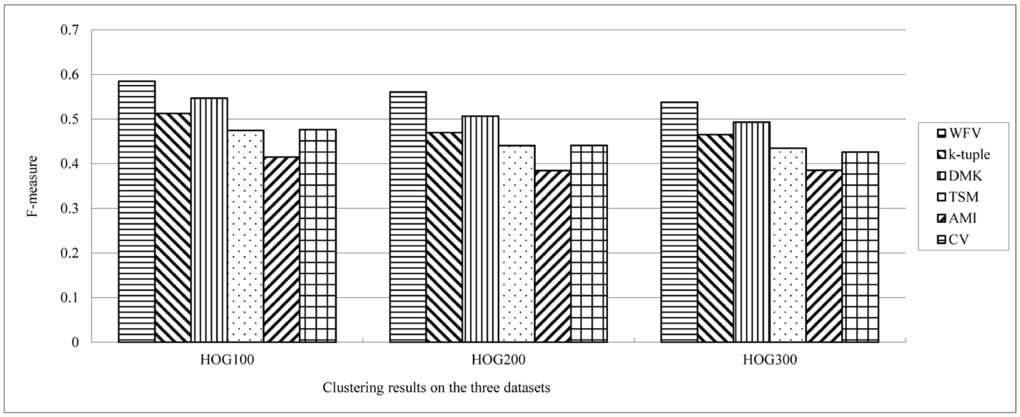

Figure 2. Clustering results in F-measure of the six alignment-free models on different datasets.

Running time

The WFV model not only improves the accuracy of clustering, but also shortens the running time. Table 2 lists the average running time of the six models on the three datasets. The running environment was as follows.

CPU: Intel Core i7 (3.40 GHz), RAM: 32 GB, OS: Windows 7 (64 bit professional edition).

It should be noted that the feature vector was built once while the K-means procedure was executed 200 times. The WFV's feature vector building time was a little longer than that of k-tuple, but it was much shorter than that of the other models. It is more important that the clustering time of WFV was the shortest. In general, the total running time of WFV was the shortest, so WFV is more suitable for enormous quantities of data.

Table 2. Running time in seconds of the alignment-free models on different datasets.

Dataset	Model	Length of feature vector	Time of building feature vector	Time of K-means clustering	Total running time
HOG100	WFV	32	8.4857	102.2634	110.7491
	k-tuple	64	7.9544	763.5223	771.4767
	DMK	64	30.2172	1550.4054	1580.6226
	TSM	12	32.7890	576.7237	609.5127
	AMI	4	167.8232	326.0293	493.8525
	CV	64	24.5168	2715.1169	2739.6337
HOG200	WFV	32	20.5239	645.8376	666.3615
	k-tuple	64	19.6306	3010.5426	3030.1732
	DMK	64	74.3083	7210.1629	7284.4712
	TSM	12	82.2772	2144.6954	2226.9726
	AMI	4	410.8531	1059.0261	1469.8792
	CV	64	60.3402	9247.6313	9307.9715
HOG300	WFV	32	24.1259	1349.6459	1373.7718
	k-tuple	64	22.5723	5560.3319	5582.9042
	DMK	64	85.1456	13785.8160	13870.9616
	TSM	12	92.8571	3329.9724	3422.8295
	AMI	4	471.9211	1577.3361	2049.2572
	CV	64	69.3221	16609.7183	16679.0404

CONCLUSIONS

This paper proposed an alignment-free model called the WFV model, which converts DNA sequences into digital sequences and uses DWT to extract features in both the time and frequency domains. We introduced an adaptive method to build DWT feature vectors with a fixed dimension. The preferred length of feature vectors for the WFV model is 32. According to the K-means clustering experimental results for several large-scale DNA datasets, WFV is superior to other models, including k-tuple, DMk, TSM, AMI, and CV. It is not only faster but also more accurate than the other models, so it is more suitable for large-scale DNA data.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (Grant #60903123, #60933003) and the Fundamental Research Funds for the Central Universities.

REFERENCES

Abbasi O and Rasi J (2011). Exonic regions finding on DNA sequences using RLS algorithm and de noising with discrete wavelet. 2011 International Symposium on Artificial Intelligence and Signal Processing (AISP). IEEE, 66-70.

Alexandrov T, Decker J, Mertens B, Deelder AM, et al. (2009). Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics* 25: 643-649.

Bauer M, Schuster SM and Sayood K (2008). The average mutual information profile as a genomic signature. *BMC Bioinformatics* 9: 48.

Bonham-Carter O, Steele J and Bastola D (2014). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief. Bioinform.* 15: 890-905.

Chang G, Wang H and Zhang T (2014). A novel alignment-free method for whole genome analysis: Application to HIV-1 subtyping and HEV genotyping. *Information Sciences* 279: 776-784.

Du P, Kibbe WA and Lin SM (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22: 2059-2065.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95: 14863-14868.

Grabherr MG, Russell P, Meyer M, Mauceli E, et al. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* 26: 1145-1151.

Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, et al. (2010). Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J. Mol. Biol.* 397: 119-143.

Heyne S, Costa F, Rose D and Backofen R (2012). GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* 28: i224-i232.

Leimeister CA and Morgenstern B (2014). Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* 30: 2000-2008.

Li KB, Issac P and Krishnan A (2004). Predicting allergenic proteins using wavelet transform. *Bioinformatics* 20: 2572-2578.

Liò P (2003). Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* 19: 2-9.

Liu L, Ho YK and Yau S (2006). Clustering DNA sequences by feature vectors. *Mol. Phylogenet. Evol.* 41: 64-69.

Liu Y (2007). Wavelet feature selection for microarray data. Proceedings of the 2007 IEEE/NIH Life Science Systems and Applications Workshop. IEEE, 205-208.

Machado JA, Costa AC and Quelhas MD (2011). Wavelet analysis of human DNA. *Genomics* 98: 155-163.

Nanni L and Lumini A (2011). Wavelet selection for disease classification by DNA microarray data. *Expert Syst. Appl.* 38: 990-995.

Nanni L, Brahnam S and Lumini A (2012). Combining multiple approaches for gene microarray classification. *Bioinformatics* 28: 1151-1157.

Pique-Regi R, Tsau ES, Ortega A, Seeger R, et al. (2007). Wavelet footprints and sparse Bayesian learning for DNA copy number change analysis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 1: I-353-I-356.

Pyatkov MI and Pankratov AN (2014). SBARS: fast creation of dotplots for DNA sequences on different scales using GA-,GC-content. *Bioinformatics* 30: 1765-1766.

Qi J, Wang B and Hao BI (2004). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.* 58: 1-11.

Rashid S and Maruf GM (2011). An adaptive feature reduction algorithm for cancer classification using wavelet decomposition of serum proteomic and DNA microarray data. Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE, 305-312.

Sharma D, Issac B, Raghava GP and Ramaswamy R (2004). Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* 20: 1405-1412.

Shi L and Huang H (2012). DNA sequences analysis based on classifications of nucleotide bases. *Springer Berlin Heidelberg* 137: 379-384.

Vinga S and Almeida J (2003). Alignment-free sequence comparison-a review. *Bioinformatics* 19: 513-523.

Wang P, Yang P, Arthur J and Yang JY (2010). A dynamic wavelet-based algorithm for pre-processing tandem mass spectrometry data. *Bioinformatics* 26: 2242-2249.

Wei D and Jiang Q (2010). A DNA sequence distance measure approach for phylogenetic tree construction. Proceedings of the 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA). IEEE, 204-212.

Yi G, Sze SH and Thon MR (2007). Identifying clusters of functionally related genes in genomes. *Bioinformatics* 23: 1053-1060.