Definition and Usage of Texture Feature for Biological Sequence

Weiyang Chen and Weiwei Li

Abstract—In recent years, sequencing technology has developed rapidly. This produces a large number of biological sequence data. Because of its importance, there have been many studies on biological sequences. However, there is still a lack of an effective quantitative method for defining and calculating texture features of biological sequences. Texture is an important visual feature. It is generally used to describe the spatial arrangement of intensities of images. Here we defined the texture features of biological sequence. Combining the digital coding of biological sequence with the calculation method of image texture features, we defined the texture features of biological sequence and designed the calculation method. We applied this method to DNA sequence features quantification and analysis. Using these quantified features, we can compute the similarity distance matrix of DNA sequences and construct the phylogenetic relationships based on the clustering of the quantified features. This method can be applied to analyze any biological sequence, and all biological sequences can be digitally coded and texture features can be calculated by this method. This is a novel study of biological sequence texture features. This will usher in a new era of quantitative and mathematical calculation of biological sequence features.

Index Terms—Biological sequence, co-occurrence matrix, texture feature, quantitative analysis

1 Introduction

In recent years, sequencing technology has developed rapidly and the next-generation sequencing technology has also been widely used [1]. This produces a large number of biological sequence data. Biological sequence data have been applied in many fields, such as the biological sequences similarity analysis, phylogenetic relationships construction, gene function analysis and protein structure analysis.

The biological sequences data can be used to categorize microRNAs [2]. The used *k*-mers are short fragments of biological sequence which length is *k*. Biological sequences analysis is useful for pre-miRNA detection, target prediction and MicroRNA categorization.

Sequences similarity analysis is an important topic in bioinformatics [3]. It is the foundation of the evolutionary relationship prediction among species, gene function prediction and protein structure prediction. In general, sequence alignment is the main method of sequences similarity analysis [4], [5]. Biological sequence alignment and analysis are the basis of drug design, constructing evolutionary tree of species, identifying the functionalities and structures [6].

With the development of high-throughput sequencing methods, more and more biological sequences data have been generated. Therefore, there is a huge demand for the computational analysis methods of biological sequences. There are many studies on biological sequences. However, there is a lack of an effective quantitative method for defining and calculating quantitative features of biological sequences. Texture is an important visual feature. It is generally used to describe the spatial arrangement of intensities of images. Here we defined the texture features of DNA sequence. Combining the digital coding of biological sequence with the calculation method of image texture features, we defined the texture features of biological sequence and designed the calculation method. Here we applied this method to DNA sequence features quantification and analysis. Using this method, we can compute the similarity distance matrix and construct the phylogenetic relationships based on the clustering of the quantified features. This method can be applied to analyze any biological sequence, and all biological sequences can be digitally coded and texture features can be calculated by this method. This is a novel study of biological sequence texture features. This will usher in a new era of quantitative and mathematical calculation of biological sequence features.

• *W. Chen and W. Li are with the School of computer science and technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong 250353, China.*

Manuscript received 1 Nov. 2019; revised 4 Feb. 2020; accepted 7 Feb. 2020. Date of publication 11 Feb. 2020; date of current version 1 Apr. 2021.
(Corresponding author: Weiyang Chen.)
Digital Object Identifier no. 10.1109/TCBB.2020.2973084

2 Method

The pipeline of our method is shown in Fig. 1. First, the DNA sequence was translated into integer vector. Second, the integer co-occurrence matrix was computed based on the integer vector. Then, the texture features can be quantified based on the co-occurrence matrix. Compared to previous studies, here we defined a simple and useful digital coding method of DNA sequence. We transfer four bases to four integers, and these integers can be used to quantify digital features.

The general pipeline is shown in Fig. 1. The DNA sequence is translated into digital vector. For example the digital coding vector of sequence ACGTAC is (1, 2, 3, 4, 1, 2). And then the co-occurrence matrix (the right panel of Fig. 1) is computed based on this digital vector. The co-occurrence matrix is computed by the co-occurrence number of adjacent values, and it should be a symmetric matrix when specifying the symmetric parameter as true.

There is an offset parameter when computing the co-occurrence matrix. It specifies the distance between the letter and its neighbor in a sequence. The 2-mers method [2] can be regarded as a special case of this method. A 2-mer over the alphabet {A, C, G, T} can generate 16 words. They are AA, AT, AG, AC, CC, CT, CA, CG, GG, GT, GA, GC, TT, TC, TA and TG. The frequency of each 2-mer can computed from each biological sequence. It is exactly the result of specifying the offset as 1 when computing the co-occurrence matrix in our method.

The defined co-occurrence matrix of biological sequence is similar to the widely used Gray Level Co-Occurrence Matrix in image processing field. The defined features of biological sequence are also similar to the image texture features [7], [8], [9], [10]. All the used texture features, computed from co-occurrence matrix, are defined as follows.

$$Entropy = -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,\ j)Ln(p(i,j)), \tag{1}$$

where L is 4, it is similar to the number of gray intensity levels in the computing of image texture features, and here it is 4, represents that the coding range of four bases is from 1 to 4. The $p(i,j)$ is the (i,j)th element of the normalized co-occurrence matrix, it is the probability of value *i* is adjacent to value *j* in the digital vector. The normalized co-occurrence matrix is obtained by dividing each element of the co-occurrence matrix by the sum of the co-occurrence matrix. After normalization, the sum of all elements of the matrix is equal to 1. These parameters are the same meaning in the following formulas.

$$Contrast = \sum_{i=1}^{L}\sum_{j=1}^{L} (i-j)^2 p(i,\ j) \tag{2}$$

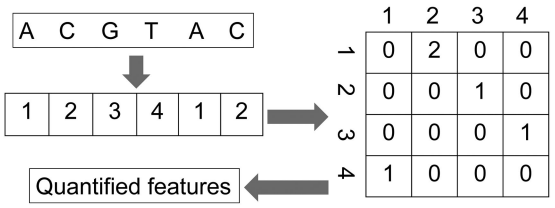

Fig. 1. The pipeline of texture features quantification of biological sequences.

$$Energy = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2 \tag{3}$$

$$Correlation = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(\frac{(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x\sigma_y}\right), \tag{4}$$

where μ_x and μ_y are the means, σ_x and σ_y are the standard deviations of p_x and p_y, the partial probability density functions.

$$Homogeneity = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(\frac{p(i,j)}{1+|i-j|}\right) \tag{5}$$

$$Autocorrelation = \sum_{i=1}^{L}\sum_{j=1}^{L}(ij)p(i,\ j) \tag{6}$$

$$Cluster\ Prominence = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(i+j-\mu_x-\mu_y\right)^4 p(i,\ j) \tag{7}$$

$$Cluster\ Shade = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(i+j-\mu_x-\mu_y\right)^3 p(i,\ j) \tag{8}$$

$$Dissimilarity = \sum_{i=1}^{L}\sum_{j=1}^{L}|i-j|p(i,\ j) \tag{9}$$

$$Maximum\ probability = \ MAX_{i,j}(p(i,\ j)) \tag{10}$$

$$Sum\ of\ squares = \sum_{i=1}^{L}\sum_{j=1}^{L}(i-\mu)^2\ p(i,\ j), \tag{11}$$

where μ is the mean of the co-occurrence matrix.

$$Sum\ average = \sum_{k=2}^{2L} k\left(\sum_{\substack{i=1\\ i+j=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) \tag{12}$$

$$Sum\ entropy = -\sum_{k=2}^{2L}\left(\sum_{\substack{i=1\\ i+j=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) Ln\left(\sum_{\substack{i=1\\ i+j=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) \tag{13}$$

$$Sum\ variance = \sum_{k=2}^{2L}(k - Sum\ entropy)^2\left(\sum_{\substack{i=1\\ i+j=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) \tag{14}$$

$$Difference\ variance = \sum_{k=0}^{L-1} k^2\left(\sum_{\substack{i=1\\ |i-j|=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) \tag{15}$$

$$Difference\ entropy = -\sum_{k=0}^{L-1}\left(\sum_{\substack{i=1\\ |i-j|=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) Ln\left(\sum_{\substack{i=1\\ |i-j|=k}}^{L}\sum_{j=1}^{L} p(i,\ j)\right) \tag{16}$$

$$Information\ measure\ of\ correlation = \frac{Entropy - HXY1}{\max(HX, HY)}, \tag{17}$$

where HX, HY are the entropies of p_x and p_y, the $HXY1$ is defined as follows,

$$HXY1 = \ -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,\ j)Ln\big(p_x(i)p_y(j)\big) \tag{18}$$

$$Inverse\ difference\ normalized = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{p(i,j)}{1+|i-j|/L} \tag{19}$$

$$Inverse\ difference\ moment\ normalized = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{p(i,j)}{1+(i-j)^2/L^2}. \tag{20}$$

For example, given a short sequence (ATAGACTCTGCTAGAGG), the co-occurrence matrix of this sequence is [0 1 5 3; 1 0 1 4; 5 1 2 1; 3 4 1 0] when specifying the symmetric parameter as true. Then this matrix can be used to compute the nineteen features. For this short sequence, the computed values of these nineteen texture features, Autocorrelation, Contrast, Correlation, ClusterP, ClusterS, Dissimilarity, Energy, Entropy, Homogeneity, MaximumP, SumSquares, SumAverage, SumVariance, SumEntropy, DifferenceVariance, DifferenceEntropy, Information1, InverseINN, and Inverse3, are 5.5, 4.125, -0.571429, 2.625, 0, 1.875, 0.107422, 2.366901, 0.390625, 0.15625, 7.253906, 5, 13.940566, 1.420116, 4.125, 1.12467, -0.277341, 0.694643, and 0.808971, respectively.

The program is freely downloadable at https://github.com/weiyangc/SequenceFeatures.

3 Results

We used the DNA sequences of twelve primate species to do the digital coding and features quantification analysis. The selected DNA sequences are widely used in previous studies [10], [11], [12], [13], [14]. Every sequence was digital coded and quantified features independently. Nineteen features were quantified for each biological sequence. These features are important texture features [7], [8], [9], [10]. These texture features include Contrast, Energy, Correlation, Homogeneity, Entropy, Cluster Shade (ClusterS), Cluster Prominence (ClusterP), Dissimilarity, Autocorrelation, Sum of squares (SumSquares), Sum average (SumAverage), Sum variance (SumVariance), Sum entropy (SumEntropy), Maximum probability (MaximumP), Information measure of correlation (Information1), Difference variance, Difference entropy, Inverse difference moment normalized (Inverse3) and Inverse difference normalized (InverseINN).

After the quantified features are obtained, the statistical analysis can be carried out. Fig. 2 shows the result of principle component analysis (PCA) based on the quantified features. This figure shows the results of PCA for twelve primate species, each species has

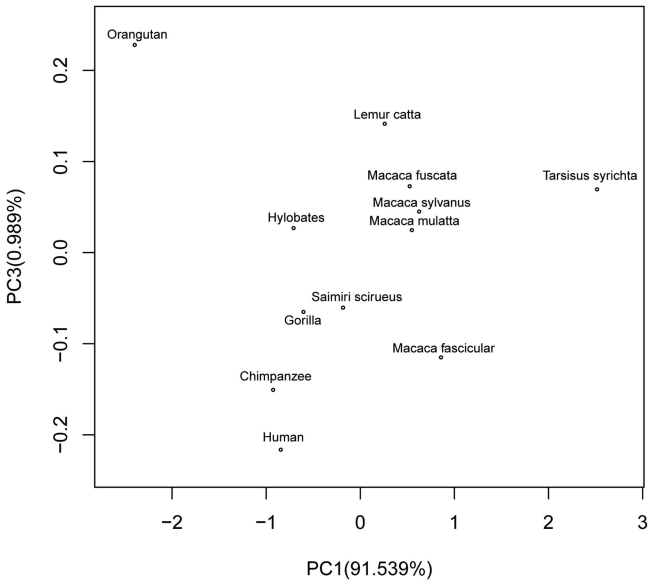

Fig. 2. The PCA analysis of twelve primate species based on the quantified features.

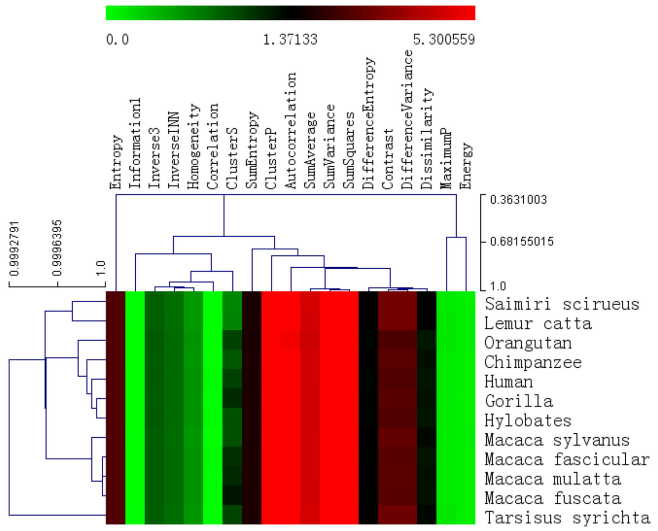

Fig. 4. The clustering analysis based on the quantified features.

nineteen features and these features were used to extract several main components. The R package was used for principle component analysis. In the figure, we can see that Macaca fuscata, Macaca sylvanus and Macaca multta have similar values in the first principal component.

Fig. 3 shows the results of PCA for the nineteen features, each feature has twelve values related to twelve primate species and these values were used to extract several main components. The R package was used for principle component analysis. This figure shows the relationship of different features.

Clustering method can classify the samples into different clusters, and there are similar pattern of the samples in the same cluster. Here we use the clustering analysis to classify both twelve primate species and nineteen features. The clustering based on the quantified features and the heat map visualization are done through the Multiple Experiment Viewer (MeV 4.9.0) [15].

Fig. 4 shows the clustering results. Each column represents one feature, and each row represents one primate species. In the clustering results of columns, we can see that the features with similar values are clustered together. For example, Difference variance and Contrast are clustered together, Sum of squares and Sum variance are clustered together, Homogeneity and Inverse difference normalized are clustered together. This means that their values are very similar. In the clustering results of rows, we can see that the clustering results of twelve primate species based on the values of these nineteen features. The results show that Macaca fuscata, Macaca multta, Macaca fascicular and Macaca sylvanus are clustered together. And human is closer to Chimpanzee and Gorilla. The results are in good agreement with previous studies [11], [13], [16].

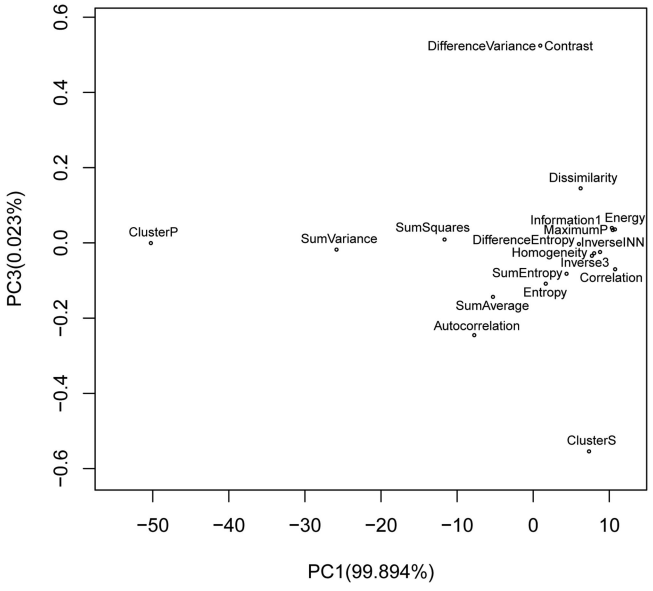

Fig. 3. The PCA analysis of the quantified nineteen features.

4 Discussions and Conclusion

In the field of biological sequence processing, many researches have been done before, such as the numerical representation [3], graphical representation [17] and similarity analysis [14], [18]. However, there is still a lack of a systematic quantitative method for defining and calculating features of biological sequences. Texture is an important visual feature. It is widely used in image processing [7], [8], [19], [20]. Here we applied the image texture features quantification theory to biological sequence analysis, and defined the texture features of biological sequence. This is useful for the alignment-free sequence similarity analysis, gene function analysis, constructing phylogenetic trees, protein structure prediction, and biological sequence retrieving in database.

The advantages of this method are as follows. First, it defined and quantified many texture features for biological sequences, which have not been proposed for biological sequences analysis before. Second, some other methods, such as the 2-mers, can be included in this method. Compared with the 2-mers method, we have made a step forward. In the 2-mers method, the frequency of each 2-mer can be computed. In our method, we can not only calculate various co-occurrence frequencies, but also define various texture features and corresponding calculation methods. Third, this method can be extended to analyze any biological sequence, and all biological sequences can be digitally coded and texture features can be calculated by this method.

Acknowledgments

This work was supported by Natural Science Foundation of Shandong Province, China (ZR2017BF041).

REFERENCES

[1] L., Liu *et al.*, "Comparison of next-generation sequencing systems," *J. Biomed. Biotechnol.*, 2012. 2012, Art. no. 251364.

[2] M. Yousef *et al.*, "MicroRNA categorization using sequence motifs and k-mers," *BMC Bioinformatics*, vol. 18, no. 1, 2017, Art. no. 170.

[3] W. Chen *et al.*, "An improved binary representation of DNA sequences and its applications," *MATCH Commun. Math, Comput. Chem.*, vol. 61, no. 3, pp. 767–780, 2009.

[4] W. Chen *et al.*, "An ant colony pairwise alignment based on the dot plots," *J. Comput. Chem.*, vol. 30, no. 1, pp. 93–97, 2009.

[5] W. Chen *et al.*, "Multiple sequence alignment algorithm based on a dispersion graph and ant colony algorithm," *J. Comput. Chem.*, vol. 30, no. 13, pp. 2031–2038, 2009.

[6] K. D. Nguyen and Y. Pan, "A knowledge-based multiple-sequence alignment algorithm," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 4, pp. 884–896, Jul./Aug. 2013.

[7] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man, Cybern.*. 3, no. 6, pp. 610–621, Nov. 1973.

[8] L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 780–795, Mar. 1999.

[9] D. A., Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Can. J. Remote Sens.*, vol. 28, no. 1, pp. 45–62, 2002.

[10] W. Chen, B. Liao, and W. Li, "Use of image texture analysis to find DNA sequence similarities," *J. Theor. Biol.*, vol. 455, pp. 1–6, 2018.

[11] K. Hayasaka, T. Gojobori, and S. Horai, "Molecular phylogeny and evolution of primate mitochondrial DNA," *Mol. Biol. Evol.*, vol. 5, no. 6, pp. 626–644, 1988.

[12] Y., Liu and Y. Zhang, "New invariant of DNA sequences based on a new matrix representation," *Combinatorial Chem. High Throughput Screen*, vol. 14, no. 1, pp. 61–71, 2011.

[13] X. Qi *et al.*, "A novel model for DNA sequence similarity analysis based on graph theory," *Evol. Bioinf. Online*, vol. 7, pp. 149–158, 2011.

[14] Y. Zhang, "A simple method to construct the similarity matrices of DNA sequence," *MATCH Commun. Math. Comput. Chem.*, vol. 60, no. 2, pp. 313–324, 2008.

[15] A. I. Saeed *et al.*, "TM4: A free, open-source system for microarray data management and analysis," *Biotechniques*, vol. 34, no. 2, pp. 374–378, 2003.

[16] M. Goodman *et al.*, "Primate evolution at the DNA level and a classification of hominoids," *J. Mol. Evol.*, vol. 30, no. 3, pp. 260–266, 1990.

[17] B. Liao and K. Ding, "Graphical approach to analyzing DNA sequences," *J. Comput. Chem.*, vol. 26, no. 14, pp. 1519–1523, 2005.

[18] B. E. Blaisdell, "A measure of the similarity of sets of sequences not requiring sequence alignment," *Proc. Nat. Acad. Sci. USA*, vol. 83, no. 14, pp. 5155–5159, 1986.

[19] W., Gomez, W. C. Pereira, and A. F. Infantosi, "Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound," *IEEE Trans. Med. Imag.*, vol. 31, no. 10, pp. 1889–1899, Oct. 2012.

[20] J. V. Marcos *et al.*, "Automated pollen identification using microscopic imaging and texture analysis," *Micron*, vol. 68, pp. 36–46, 2015.