Soft Computing (2019) 23:9175–9188
https://doi.org/10.1007/s00500-018-3602-2

FOCUS

Nucleosome positioning based on generalized relative entropy

Mengye Lu[1,2] · **Shuai Liu**[1,2]

Published online: 30 October 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract
Nucleosome positioning played significant roles in various biological processes. With the development of high-throughput techniques, many methods and software were developed for nucleosome positioning. Although results with high accuracy (Acc) were obtained, the key factors for determining nucleosome positioning under less time complexity remain unresolved. Therefore, combining generalized relative entropy with self-similarity of DNA sequences, a novel method of nucleosome positioning was proposed for predicting nucleosome positioning in human, worm, fly and yeast genomes, respectively. Experimental results showed that prediction Acc of nucleosome positioning in aforementioned datasets reached 87.78%, 87.98%, 83.36% and 100%, respectively. Furthermore, it was found that five-nucleotide and six-nucleotide sequences were the determinant factors in nucleosome positioning.

Keywords Nucleosome positioning · Generalized relative entropy · Random forest · Support vector machines

1 Introduction

Nucleosomes are the basic unit of eukaryotic chromatin, and each one is constructed by a histone octamer that wrapped tightly by a DNA sequence with 147 base pair (bp). Adjacent nucleosomes are connected by linker DNA. As shown in Fig. 1, the histone octamer is constructed by two-molecular H2A, H2B, H3 and H4. Due to the influence of H1 histone, nucleosomes form a stable structure. Position of nucleosome is related to various biological processes, such as DNA replication and RNA splicing (Yasuda et al. 2005; Berbenetz et al. 2010). Besides, position of nucleosome is always dynamically changing in these processes. Therefore, nucleosome positioning is necessary to have an in-depth understanding about biological processes.

Earlier, Satchwell et al. (1986) found a 10-bp interval repetition of AA/TT/TA that appeared in the region of core DNA. Besides, some sequences of the core DNA appeared periodically (Ioshikhes et al. 1996, 2006). Afterward, it was found that nucleosome deficiency appeared in poly (dA:dT) fragments (Segal and Widom 2009).The above findings were considered as important nucleosome positioning signal and indicated that nucleosome positioning was sequence dependent to some extent.

With the development of high-throughput techniques, high-resolution nucleosome positioning maps of many species have been obtained (Lee et al. 2007; Schones et al. 2008). Therefore, various methods and tools of nucleosome positioning were proposed. Many predictors were constructed based on frequencies information of nucleotide sequences combinations. Ioshikhes et al. (2006) analyzed characteristics of TATA box that occurred in core DNA and linker DNA, respectively and then applied these characteristics to nucleosome positioning. Peckham et al. (2007) predicted position of nucleosome based on the characteristics of gene sequences in promoter regions. Afterward, Kaplan et al. (2009) analyzed DNA encode of nucleosome organization in eukaryotic genome and then predicted position of nucleosome based on it. Afterward, Xi et al. (2010) used hidden markov model (HMM) in nucleosome positioning. Polishko et al. (2012) applied a modified Gaussian mixture model to nucleosome positioning. Struhl and Segal (2013) predicted position of nucleosome based on characteristics of gene sequences. Besides, Freeman et al. (2014) employed molecular models of DNA and proteins to elucidate vari-

Communicated by A. K. Sangaiah, H. Pham, M.-Y. Chen, H. Lu, F. Mercaldo.

✉ Shuai Liu
cs_liushuai@imu.edu.cn

[1] College of Computer Science, Inner Mongolia University, Hohhot, China

[2] Inner Mongolia Key Laboratory of Social Computing and Data Processing, Inner Mongolia University, Hohhot, China

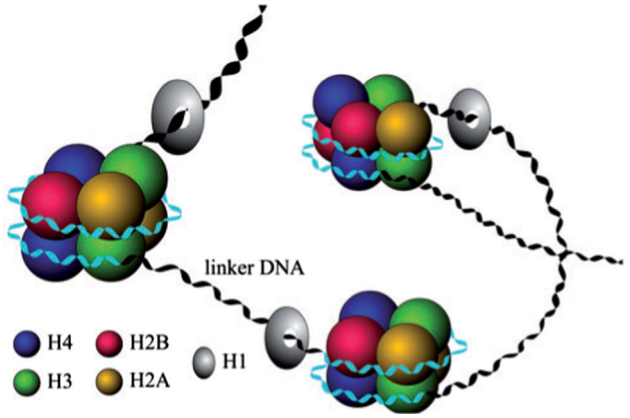

Fig. 1 Nucleosomes are the basic unit of eukaryotic chromatin, and each one is constructed by a histone octamer that wrapped tightly by a DNA sequence with 147 base pair (bp)

ous aspects of nucleosome positioning. Recently, Chen et al. (2016) have used deformation energy to analyze nucleosome positioning in saccharomyces cerevisiae genomes and Tahir and Hayat (2016) constructed a sequence predictor iNuc-STNC for nucleosome positioning. Core DNA sequences can be identified based on the above model; however, the key factors influencing nucleosome positioning remained unclear.

Afterward, Awazu (2017) constructed a linear regression model based on incorporation of frequencies and distributions for nucleotide sequences with different length, and applied it to nucleosome positioning. The position of nucleosome and the key factors influencing nucleosome positioning can be determined. However, the variables of model were chosen based on the stepwise forward selection method, which took a lot of time. Zhang et al. (2018a, b) studied nucleosome positioning using improved convolutional neural networks. Compared with other methods, Acc obtained by this method was higher than that obtained by other proposed method. However, this method was based on deep learning, which took a lot of time and required higher technical support.

Besides, combining DNA sequence information and DNA structural properties, some predictors were proposed. For example, based on the information of DNA sequences and physical structure, Chen et al. (2012) and Guo et al. (2014) proposed predictors iNuc-PhysChem and iNuc-PseKNC, respectively. Furthermore, Flores and Orozco (2011), Tolstorukov et al. (2008) and Woo et al. (2013) developed some tools for nucleosome positioning.

Information entropy is an abstract concept that is often used to measure the degree of confusion of a system. It includes relative entropy, cross-entropy and mutual information and is widely used in various fields (Zhang and Wu 2008, 2011; Yudong et al. 2015). Relative Entropy (RE) is an important method that described dissimilarity between two different probability distributions, which is widely used in various fields.

Based on RE, a new distance measure method to distinguish different gene sequences was proposed (Benson 2002; Vernikos and Parkhill 2006). Besides, Magliery and Regan (2005) applied RE to identify unconceived hypervariable positions, and Wang and Samudrala (2006) applied RE to search for conserved positions of gene sequences. Vacic et al. (2007) developed a tool for discovery and visualization differences of amino acid composition. Afterward, Astrovskaya et al. (2011) proposed a method to infer viral quasispecies spectra and utilized relative entropy to measure prediction quality. Chen and Zhou (2012) proposed a relative entropy approach in group decision making. Besides, Beigi and Gohari (2014) achieved quantum achievability proof via collision relative entropy. Gibb and Strimmer (2015) proposed an approach for identifying differentially expressed proteins using binary discriminant analysis based on relative entropy. Recently, Sarosi and Ugajin (2016) have studied the relative entropy and the trace square distance in two-dimensional conformal field theories. Shao et al. (2017) studied quantum coherence quantifiers based on α relative entropy.

Many predictors were proposed for nucleosome positioning in various species. However, the sequence predicted as core DNA sequences in one species may be predicted as linker DNA sequences in another species, which showed that function of the gene sequence in assisting or inhibiting nucleosome formation was dependent on species. Besides, it took a lot of time to find the key factors influencing nucleosome positioning. Furthermore, biological common and unique characteristics in nucleosome positioning were ignored in these methods. Therefore, GRE method was proposed to extract information of DNA sequences; meanwhile, the key factors influencing nucleosome positioning were determined based on random forest (RF).

The rest of paper was organized as follows. Section 2 introduces materials and methods, including sources of data and construction of benchmark datasets, generalized relative entropy, construction processes of prediction model and evaluation metrics of model performance; Sect. 3 introduces detailed experimental processes and analyzed experimental results; Sect. 4 summarizes the contents above.

2 Materials and methods

2.1 Benchmark datasets of core DNA and linker DNA sequences

In this paper, benchmark datasets of human, worm, fly were constructed by Guo et al. (2014) and benchmark dataset of yeast was constructed by Chen et al. (2015). In order to construct benchmark datasets, entire genome data

and nucleosome positioning data were provided. Detailedly, benchmark dataset of human was constructed by Guo et al. (2014), entire genome sequences were available at UCSC genome database (http://hgdownload.cse.ucsc.edu), where 18hg version was used for human genome. Due to huge data, chromosome 20 was extracted as entire genome data of human (Liu et al. 2011). Nucleosome positioning data were from Schones et al. (2008) Detailedly, data were available from http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx. Each of DNA fragments was assigned a nucleosome formation score to reflect its propensity to form nucleosome. The higher the score was, the more likely the fragment would be in forming a nucleosome. Thus, those fragments with the highest scores were chosen as core DNA sequences, while those with the lowest scores were chosen as linker DNA sequences. In order to eliminate the influence of redundant data on experimental results, the CD-HIT software was used to remove data with high sequences similarity and the threshold was set at 80% (Fu et al. 2012). Finally, benchmark dataset of human was obtained.

Similarly, benchmark datasets of worm and fly were constructed by Guo et al. (2014). Entire genome sequences of worm and fly were downloaded from UCSC database (http://hgdownload.cse.ucsc.edu), WS 170/ ce 4 version and BDGP Release 5 version were used for worm and fly genomes, respectively. Compared nucleosome positioning data were available at http://hgdownload.cse.ucsc.edu and http://atlas.bx.psu.edu, respectively. Detailedly, nucleosome positioning data of fly were from Mavrich et al. (2008). Using the same strategy as human, benchmark datasets of these species were obtained.

Besides, Chen et al. (2015) constructed benchmark dataset of yeast. Entire genome sequences of yeast were downloaded from http://www.yeastgenome.org/ and compared nucleosome positioning data came from Lee et al. (2007). With the same strategy, benchmark dataset of yeast was obtained.

Benchmark datasets are defined as Eq. 1

$$S_k = S_k^+ + S_k^- \tag{1}$$

where k ranges from 1 to 4, and S_1, S_2, S_3 and S_4 represent human, worm, fly and yeast benchmark datasets, respectively. Dataset S_1^+ involves 2273 core DNA sequences, and S_1^- involves 2300 linker DNA sequences; dataset S_2^+ involves 2567 core DNA sequences, and S_2^- involves 2608 linker DNA sequences; dataset S_3^+ involves 2900 core DNA sequences, and S_3^- involves 2850 linker DNA sequences; dataset S_4^+ involves 1880 core DNA sequences, and S_4^- involves 1740 linker DNA sequences. The sequence length of human, worm and fly genomes is 147 bp, and sequence length of yeast genome is 150 bp. Benchmark datasets of S_1,S_2 and S_3 are given in supplementary data of references Guo et al. (2014), and benchmark dataset of S_4 is given in supporting information of references Chen et al. (2016).

2.2 Generalized relative entropy

In probability theory, relative entropy is used to measure dissimilarity for two kinds of distributions. The smaller the relative entropy is, the more similar the two kinds of distribution are. In particular, the relative entropy of two identical distributions is zero. Relative entropy of discrete random variable is defined as Eq. 2

$$\mathrm{RE}(X, Y) = \sum_{i=1}^{s} p_x(i) \cdot \log \frac{p_x(i)}{p_y(i)} \tag{2}$$

where $p_x(i)$ and $p_y(i)$ denote two kinds of discrete probability distributions, s denotes the number of states in the state space, i is a random variable in the state space, and the value of i ranges from 1 to s.

Relative entropy is not a distance metric, and it does not a finite upper bound. To measure the differences of two vectors in the high-dimensional space, GRE is proposed. It is defined as Eq. 3

$$\begin{aligned} d(X, Y) = &\sum_{i=1}^{s} \left(p_x(i) \cdot \log \frac{k \cdot p_x(i)}{(k-1)p_x(i) + p_y(i)} \right) \\ &+ \sum_{i=1}^{s} \left(p_y(i) \cdot \log \frac{k \cdot p_y(i)}{p_x(i) + (k-1)p_y(i)} \right) \\ &+ r \cdot \log \left(1 + \frac{1}{k-1} \right)^2 \end{aligned} \tag{3}$$

where $k \geq 1, r = 0$ when $X = Y$; otherwise, $r = 1$ when $X \neq Y$. Parameter k denotes the weight of the probability of occurrence of each state. The meaning of i and s is same as that in Eq. 2.

The Markov model is a statistical model that can be used to describe the Markov process. The generation of DNA sequences can be seen as a Markov process. Assuming that the sequence that appeared in the next state depends on the previous state, here we use a first-order 4 Short form of author list Markov chain to represent the DNA sequence. Therefore, the generation probability of DNA sequence can be expressed by Eq. 4

$$p_n = p(s_0) * a_{ij}^{n-1} \tag{4}$$

where p_n represents generation probability of DNA sequence with length n, s0 represents initial state, $p(s_0)$ represents occurrence probability of initial state, and n represents sequence length of DNA. A single-nucleotide A C G T is represented by 1 2 3 4, respectively. i and j are two random variables, and the values range from 1 to 4. Therefore,

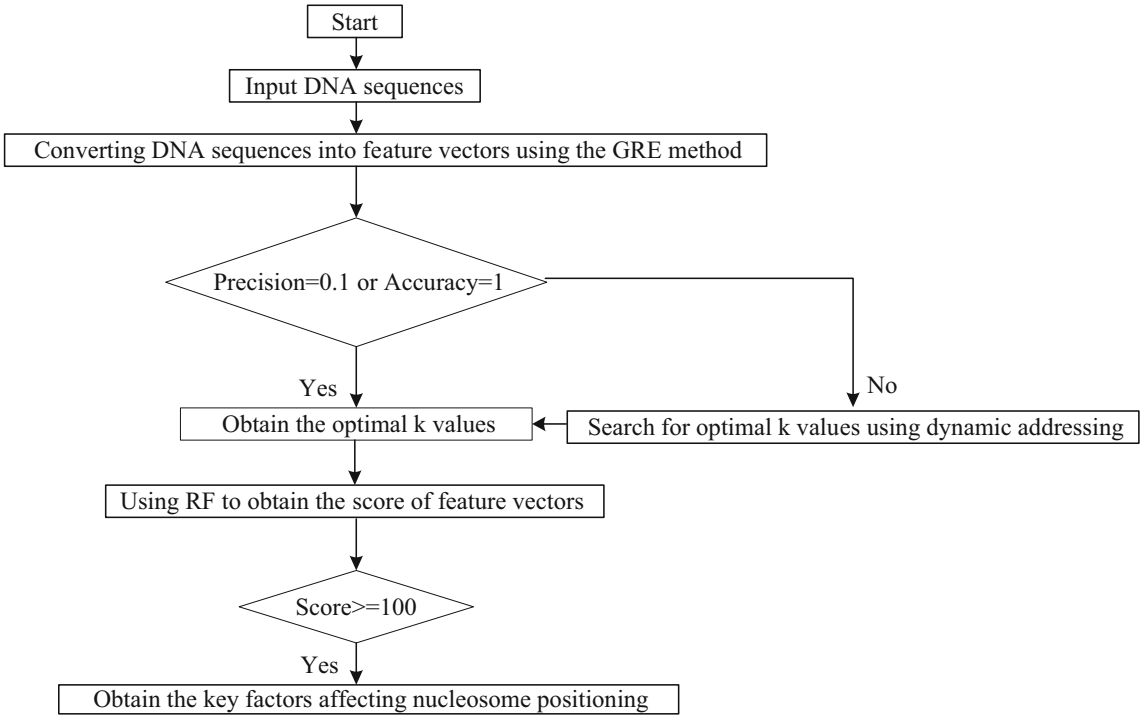

Fig. 2 Nucleosome positioning flowchart

a_{ij} represents state transition probability between different nucleotides.

Assuming that the length of the DNA sequence is m and m is far less than n, the generation probability of DNA sequence can be expressed by Eq. 5

$$p_m = t * p(s_0) * a_{ij}^{m-1} \tag{5}$$

where t belongs to Q and Q represents rational number. If DNA sequences have self-similarity, there must be a constant that satisfies Eq. 6

$$\frac{p_n}{p_m} = k \tag{6}$$

where k is a constant. Because the ratio of p_n to p_m is a_{ij}^{n-m}/t, DNA sequences satisfy self-similarity. Based on self-similarity of DNA sequences and GRE method, core DNA sequences can be identified.

2.3 Prediction model

In order to predict nucleosome positioning, generalized relative entropy (GRE) was proposed. The model was constructed by the following steps.

Step 1 Calculate all types of di-nucleotide frequency in core DNA sequences and obtain their distribution $X_2 = \{x_1, x_2, \ldots, x_{16}\}$, and x_i denotes frequency of the ith di-nucleotide. Using the same methodology, calculate all types of di-nucleotide frequency in linker DNA sequences and obtain their distribution $X_2' = \{x_1', x_2', \ldots, x_{16}'\}$.

Step 2 Similarly, calculate all types of di-nucleotide frequency in all core DNA sequences and linker DNA sequences and then obtain their distribution $Y_2 = \{y_1, y_2, \ldots, y_{16}\}$.

Step 3 Calculate generalized relative entropy of each DNA sequence according to Eq. 3, which constructs two-dimensional feature vectors $d_2 = \mathrm{GRE}(d_2^+, d_2^-)$, where $d_2^+ = d(X_2, Y_2)$ and $d_2^- = d(X_2', Y_2)$.

Step 4 Following steps 1 to 3, calculate the generalized relative entropy of other types of DNA fragments, of which trinucleotide, four-nucleotide, five-nucleotide and six-nucleotide. It constructs ten-dimensional feature vectors $d_{10} = \mathrm{GRE}(d_2^+, d_2^-, d_3^+, d_3^-, d_4^+, d_4^-, d_5^+, d_5^-, d_6^+, d_6^-)$.

Step 5 Calculate classification importance score based on random forest and search for crucial feature vectors in nucleosome positioning according to classification importance score.

Step 6 Put those feature vectors obtained in step 5 into SVM to recognize core DNA sequences.

Step 7 Use jackknife test and tenfold cross-validation to examine prediction performance of GRE method.

The detailed processing process of nucleosome positioning is shown in Fig. 2.

2.4 Machine learning algorithm

Machine learning (ML) is a technology that studies the process of making machines intelligent through learning of past experience. In recent years, due to the development of artificial intelligence (AI), machine learning algorithms have been widely used (Zhang et al. 2015; Petralia et al. 2015; Ide et al. 2016; Lin et al. 2017; Ismail et al. 2017; Karlekar and Gomathi 2018; Sinoquet 2018).

BP neural network can solve different classification problems by simulating the structure of human brain. However, because BP neural network is an optimization method of local search, the algorithm is easy to fall into the local optimal solution, which can lead to training failure (Meng et al. 2018; Zhang et al. 2017).

Support vector machine (SVM) is a method for both linear and nonlinear data classifications, and the main idea of SVM is to map nonlinear data in a low-dimensional space into a high-dimensional space and then search for the optimal hyperplane using a kernel function to separate data in one class from another class. It has been widely used in the field of bioinformatics (Bhasin and Raghava 2004; Wan et al. 2013).

Random forest (RF) is an important class cation algorithm which can separate two different types of data; meanwhile, it provides the importance scores of feature attributes for classification. It has been widely used in the field of bioinformatics (Petralia et al. 2015; Ide et al. 2016; Zhang et al. 2016; Rahman et al. 2017; Taherzadeh et al. 2017; Ismail et al. 2017; Sinoquet 2018; Fabris et al. 2018).

Therefore, in this paper, due to its powerful nonlinear mapping capabilities, SVM was selected to be a classifier to distinguish core DNA and linker DNA. Meanwhile, RF was used to search for key factors in nucleosome positioning because it can provide the importance scores of feature attributes for classification.

Besides, the software package LIBSVM 3.22 was used to be as an implementation of SVM, which was available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Radial basis function was selected as the kernel function. The parameters c and g in the training model were determined by a grid search approach.

2.5 Metrics for performance evaluation

In this paper, prediction performance of GRE model was evaluated by jackknife test and tenfold cross-validation. Detailedly, jackknife test was used to examine prediction performance of GRE model in human, worm and fly genomes. Meanwhile, tenfold cross-validation was used to examine prediction performance of GRE model in yeast genome. These methods were widely applied to evaluate prediction quality of the previously proposed model (Chen et al. 2012; Guo et al. 2014; Chen et al. 2016; Tahir and Hayat 2016; Awazu 2017).

Here, Eqs. 7–10 are defined to evaluate performance of model. Detailedly, TP is defined as the number that core DNA sequences correctly predicted core DNA sequences, FP is defined as the number that linker DNA sequences incorrectly predicted core DNA sequences, FN is defined as the number that core DNA sequences incorrectly predicted linker DNA sequences, and TN is defined as the number that linker DNA sequences correctly predicted linker DNA sequences. Using jackknife test and tenfold cross-validation, the following metrics are obtained:

$$S_n = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$S_p = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{8}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{9}$$

$$\mathrm{Mcc} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{10}$$

where S_n, S_p, Acc and Mcc represent sensitive, specificity, accuracy and Mathew's correlation coefficient, respectively.

3 Results and discussion

3.1 Parameter optimization

As shown in Eq. 3, our proposed method is based on a parameter k, where k is the weight factor and reflects distributions of core DNA sequences and linker DNA sequences in the entire genome. Furthermore, an appropriate k can reflect distributions of core DNA and linker DNA in high accuracy so that we can separate core DNA from linker DNA. Thus, searching for the optimal values of parameter k is necessary.

In order to record results of search, k_1^*,k_2^* and k_3^* were defined as the optimal parameter values obtained in the first, second and third parameter optimizations, respectively, and k^* was defined as the optimal parameter values used to construct GRE model. Due to $k \geq 1$, in order to obtain the optimal parameter values and, meanwhile, reduce computational time, the following strategy was adopted.

Step 1 Search for the optimal parameter values according to Eq. 11:

$$\begin{aligned} &2 \leq k \leq 10, \quad \text{with step } \Delta = 1 \\ &10 \leq k \leq 200, \text{ with step } \Delta = 10 \end{aligned} \tag{11}$$

Table 1 The optimal parameter values in four species

Species	k_1^*	k_2^*	k_3^*	k^*
Human	190	184	184.0	184.0
Worm	140	146	145.8	145.8
Fly	8	8.5	–	8.5
Yeast	60 70 80 90 100	–	–	60 70 80 90 100

Step 2 Based on Eq. 11, k_1^* was obtained. Then, the optimal parameter values were searched according to Eq. 12.

$$\begin{aligned} &2 \le k_1^* \le 10, \quad \text{with step } \Delta = 0.1 \\ &10 \le k_1^* \le 200, \quad \text{with step } \Delta = 1 \end{aligned} \tag{12}$$

Step 3 Based on Eq. 12, k_2^* was obtained. Then, the optimal parameter values were searched according to Eq. 13.

$$\begin{aligned} &2 \le k_2^* \le 10, \quad \text{with step } \Delta = 0 \\ &10 \le k_2^* \le 200, \text{ with step } \Delta = 0.1 \end{aligned} \tag{13}$$

In Eq. 13, step with 0 meant that we stopped search and obtained the optimal parameter values k^*.

In this paper, jackknife test was used to search for the optimal parameter values in human, worm and fly genomes, respectively. Similarly, tenfold cross-validation was used to determine the optimal parameter values in yeast genome. In our experiment, prediction accuracy of fly genome declined obviously when parameter k ranged from 110 to 200. Thus, in order to reduce computational time, optimal parameter values were searched in a range of 2 to 100. Besides, prediction accuracy of yeast genome reached 1 when parameter k was equal to 60, 70, 80, 90 and 100. Therefore, the optimal parameter values of yeast genome were obtained. As for human and worm genomes, the optimal parameter values were searched in a range of 2–200.

Detailedly, the optimal parameter values were determined by the method "shortening the interval length gradually" and precision of the optimal parameter values was set at 0.1. In the processes of searching for the optimal parameter values, search was terminated when prediction accuracy reached 1 or precision of the optimal parameter values reached 0.1. The processes of parameter optimization in human genome were as follows.

Step 1 Calculate prediction accuracy when parameter k varied from 2 to 200, respectively. Then, choose that parameter k with the highest prediction accuracy as k_1^*. As shown in Table 1, $k_1^* = 190$.

Step 2 Search for those values adjacent to k_1^* and choose smaller value as starting point of the second parameter optimization and bigger value as terminal point of the second parameter optimization. Thus, the range of the second parameter optimization was determined. Then, prediction accuracy was calculated when parameter k varied from 180 to 200 with a step of 1. Among these k *values*, the parameter k with the highest prediction accuracy was chosen k_2^*, where $k_2^* = 184$.

Step 3 Determine the range of parameter optimization using the same way as step 2. Then, search for the optimal parameter values in the scope between 183 and 185 with a step of 0.1. Compared with different results obtained by different parameters k, the parameter k with the highest prediction accuracy was chosen as k_3^*, where $k_3^* = 184.0$. The precision of the optimal parameter value reached 0.1. Hence, the parameter value used to construct GRE model was obtained, where $k^* = k_3^* = 184.0$.

Using the same strategy, the optimal parameter values of worm, fly and yeast genomes were obtained. As shown in Table 1, the optimal parameter values of four species were obtained ($k^* = 184.0, = 145.8, = 8.5$ and $= 60, = 70, = 80, = 90, = 100$ for human, worm, fly and yeast genomes, respectively). In the second parameter optimization, the precision of the optimal parameter value in fly genome reached 0.1. Therefore, the optimal parameter value in GRE model was obtained. Besides, prediction accuracy of yeast genome reached 1, when k was equal to 60, 70, 80, 90 and 100. As for yeast genome, the optimal parameter values were obtained in the first parameter optimization. The parameter optimization processes and the optimal parameter values of four species are shown in Figs. 3, 4, 5, 6 and Table 1, respectively. As shown in Table 1, k_1^*,k_2^* and k_3^* represented the optimal parameter values obtained in the first, second, third parameter optimizations, respectively. k^* was the optimal parameter value in RGE model. "-" represented null value. As for fly genome, the optimal parameter value was obtained in the second parameter optimization. Therefore, the value of k_3^* was null value. Similarly, the optimal parameter values

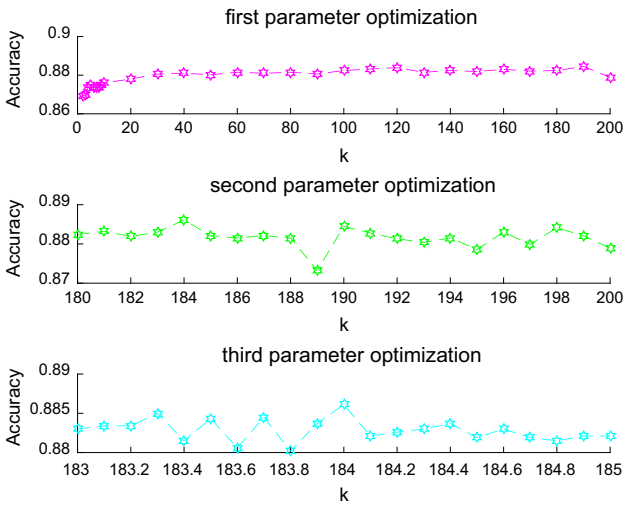

Fig. 3 Parameter optimization processes in human genome

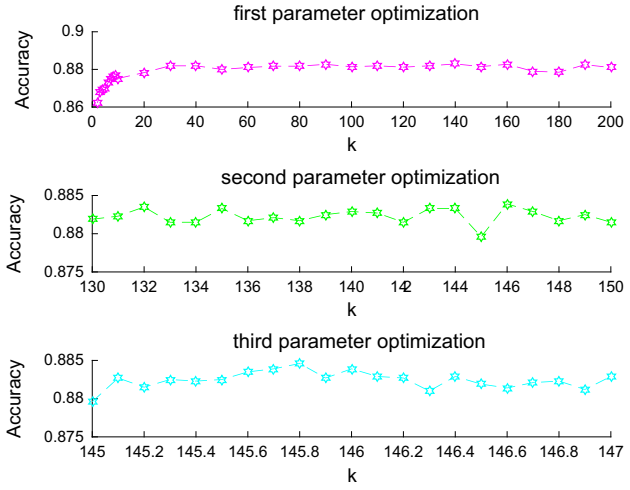

Fig. 4 Parameter optimization processes in worm genome

of yeast genome were obtained in the first parameter optimization. Therefore, both k_2^* and k_3^* were null value.

In order to analyze the effect of k values on the accuracy of SVM in different species, Table 2 is shown. In Table 2, maximum and minimum were used to represent Acc under the best case and the worst case, respectively. Meanwhile, $k_{\max}$ and $k_{\min}$ were used to represent k values corresponding to the maximum accuracy Nucleosome positioning based on generalized relative entropy 7 and the minimum accuracy, respectively. – represented null value.

In order to analyze the relation between parameter k and distribution of core DNA and linker DNA, Figs. 7, 8, 9, 10 and 11 were provided. X-axis represented DNA sequences. Detailedly, values of X-axis represented core DNA sequences when $0 \leq x \leq 1779$, and values of X-axis represented linker DNA sequences when $1880 \leq x \leq 3619$. Y-axis represented the feature values obtained by GRE model. For simplicity, 2+, 2−, 3+, 3−, 4+, 4−, 5+, 5−, 6+ and 6− in legend denoted

Fig. 5 Parameter optimization processes in fly genome

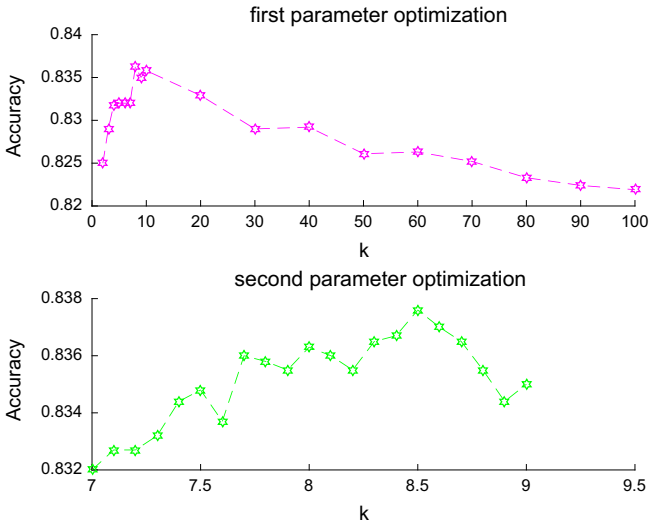

Fig. 6 Parameter optimization processes in yeast genome

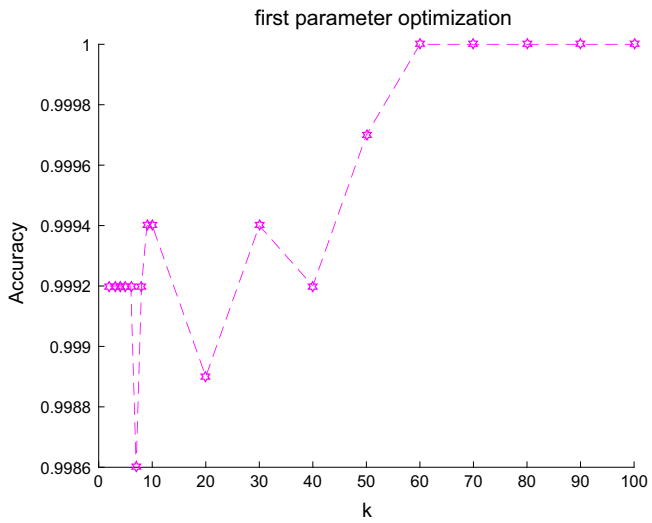

Table 2 The effect of k value on accuracy in different species

Species	k_{max}	Maximum	k_{min}	Minimum
Human	184	0.8861	2	0.8695
Worm	145.8	0.8846	2	0.8622
Fly	8.5	0.8376	100	0.8219
Yeast	60 70 80 90 100	1	7	0.9986

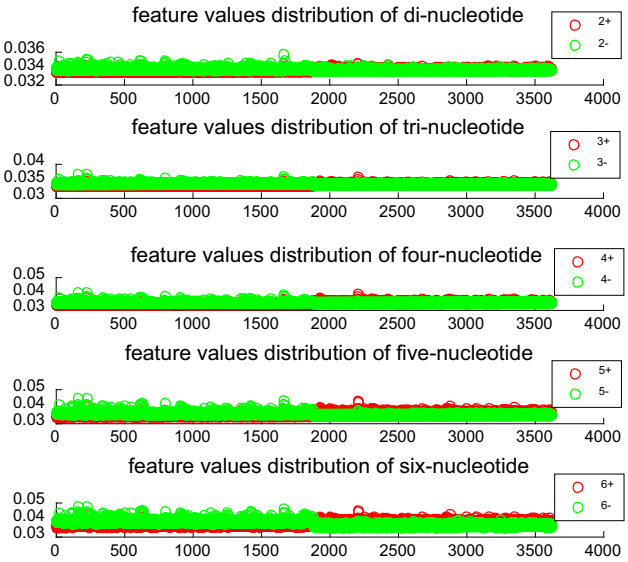

Fig. 7 Feature values distribution ($k = 60$)

positive di-nucleotide, negative di-nucleotide, positive trinucleotide, negative trinucleotide, positive four-nucleotide, negative four-nucleotide, positive five-nucleotide, negative five-nucleotide, positive six-nucleotide and negative six-nucleotide, respectively. As shown in Figs. 7, 8, 9, 10 and 11, distributions of feature values obtained by positive nucleotide sequences and negative nucleotide sequences were different in core DNA and linker DNA. Furthermore, the feature values obtained by positive nucleotide sequences were smaller than those obtained by negative nucleotide sequences in core DNA, while the feature values obtained by negative nucleotide sequences were smaller than those obtained by positive nucleotide sequences in linker DNA. Thus, distributions of feature values in core DNA and linker DNA were different. Based on these, core DNA sequences can be recognized in high accuracy. The above results indicated that prediction results obtained by GRE model were consistent with the real distributions of core DNA and linker DNA in yeast genome when the optimal parameter values were equal to 60, 70, 80, 90 and 100. Therefore, the optimal parameter values can reflect real distribution of core DNA and linker DNA to some extent.

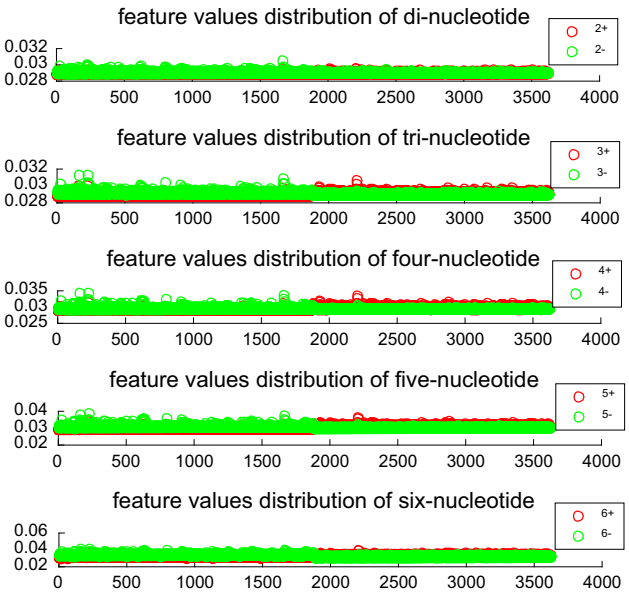

Fig. 8 Feature values distribution ($k = 70$)

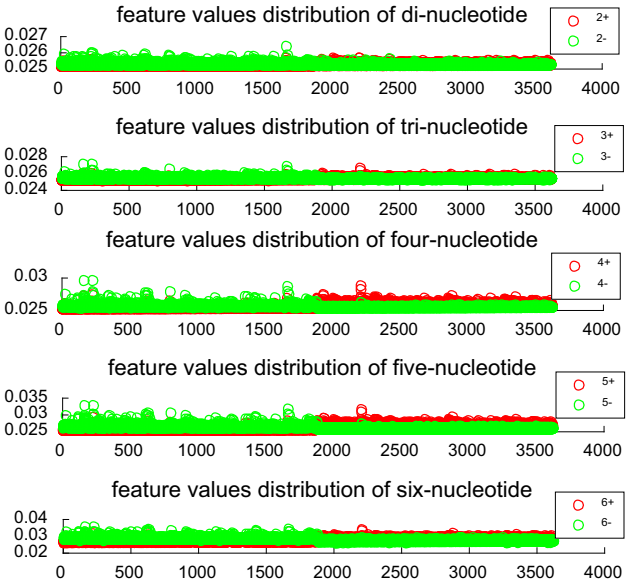

Fig. 9 Feature values distribution ($k = 80$)

3.2 Feature selection

Random forest was an important classification algorithm. It can separate two different types of data and, meanwhile, provide the importance of feature attributes for classification. Therefore, random forest was used to find the crucial factors in nucleosome positioning. First, two parameters needed to be set, including the number of trees in forest and the number of node split attributes. In this paper, the number of trees in forest was set as 50, 100,150, 200, 250, 300, 350, 400, 450 and 500, respectively, because ten-dimensional feature vectors were obtained based on GRE method and the number of

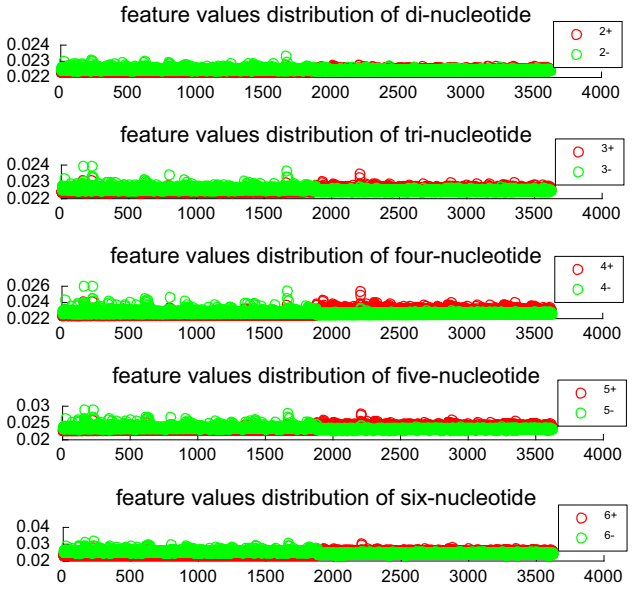

Fig. 10 Feature values distribution ($k = 90$)

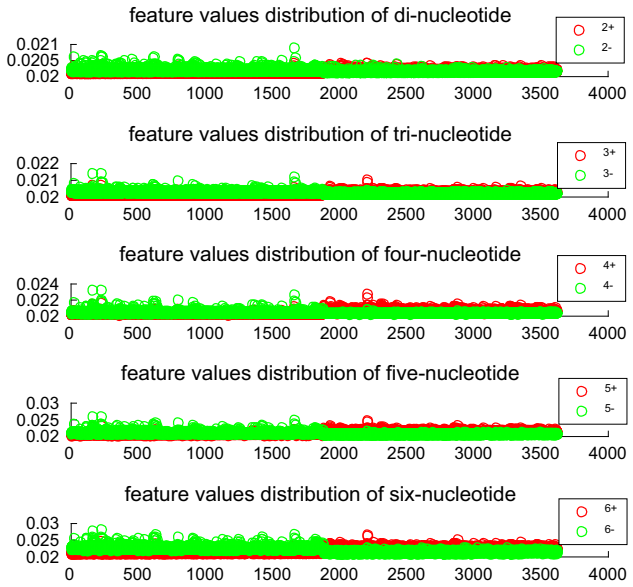

Fig. 11 Feature values distribution ($k = 100$)

node split attributes was generally set as root mean square of the number of classification attribute. Therefore, the number of node split attributes was set as 3. Next, under the optimal parameters of GRE model, calculate the importance score of ten feature vectors when the number of trees in forest was set as 50, 100, 150, 200, 250, 300, 350, 400, 450 and 500, respectively. The feature vectors whose score exceeded 100 are shown in Table 3. In Table 3, v1, v2, v3, v4, v5, v6, v7, v8, v9 and v10 presented those feature vectors obtained by positive di-nucleotide, negative di-nucleotide, positive trinucleotide, negative trinucleotide, positive four-nucleotide, negative four nucleotides, positive five-nucleotide, negative five-nucleotide, positive six-nucleotide and negative six-nucleotide, respectively. Then, set threshold of score as 100 to search for those feature vectors whose score threshold exceeded 100. Finally, search for the common feature vectors whose score threshold exceeded 100 in four species. Those feature vectors were considered as the key factors affecting nucleosome positioning. As shown in Table 3, v10, v9, v8, v5 and v7 were the common feature vectors of four species. Therefore, positive four-nucleotide, positive five-nucleotide, negative five-nucleotide, positive six-nucleotide and negative six-nucleotide sequences were considered as the key factors affecting nucleosome positioning.

Table 3 The optimal parameter values in four species

Species	Value of k	Feature vectors
Human	190	v10, v9, v5, v7, v8, v3, v6
Worm	140	v10, v8, v4, v9, v6, v7, v5, v3, v1
Fly	8	v10, v8, v9, v7, v5, v6, v4, v3, v2, v1
Yeast	60	v10, v9, v8, v5, v7
Yeast	70	v10, v9, v8, v5, v7
Yeast	80	v10, v9, v8, v5, v7
Yeast	90	v10, v9, v8, v5, v7
Yeast	100	v10, v9, v8, v5, v7

In order to analyze the effect of the number of trees on Acc in different species, Table 4 was provided. As shown in Table 4, maximum and minimum were used to represent Acc under the best case and the worst case, respectively. Meanwhile, $\text{tree}_{\max}$ and $\text{tree}_{\min}$ were used to represent the number of trees corresponding to the maximum accuracy and the minimum accuracy. Besides, - represented null value.

3.3 Prediction results for human, worm and fly genomes

Using jackknife test, S_n, S_p, Acc and Mcc were obtained. Acc (= 0.8778, = 0.8798 and = 0.8336 for human, worm and fly genomes, respectively) obtained by our proposed model was higher than those obtained by iNuc-PseKNC (Guo et al. 2014) for human, worm and fly genomes, higher than those obtained by iNuc-PseSTNC (Tahir and Hayat 2016) for human and fly genomes, and higher than those obtained by 3LS model (Awazu 2017) for worm genomes (Tables 5, 6, 7). Compared with those methods provided by Guo and Tahir, our proposed method can find the key factors influencing nucleosome positioning with lower time complexity. Although the method proposed by Awazu can find key factors influencing nucleosome positioning, its time complexity

Table 4 The effect of the number of trees on accuracy in different species

Species	$tree_{max}$	Maximum	$tree_{min}$	Minimum
Human	100	0.8662	200 400	0.8596
Worm	500	0.8611	50	0.8554
Fly	500	0.7847	400	0.78
Yeast 60	100 150 200 300 350 400 450 500	1	50 250	0.9997
Yeast 70	50 100 150 200 300 350 400 450 500	1	–	–
Yeast 80	50 100 150 200 300 400 450 500	1	200 350	0.9997
Yeast 90	100 150 200 300 350 400 450 500	1	–	–
Yeast 100	100 150 200 300 350 400 450 500	1	–	–

Table 5 The prediction performance for human genome

Metrics	Methods			
	GRE	3LS	iNuc-PseKNC	iNuc-PseSTNC
Acc	0.8778	0.9001	0.8627	0.8760
S_n	0.9107	0.9169	0.8786	0.8931
S_p	0.8452	0.8835	0.8470	0.8591
Mcct	0.7573	0.8006	0.73	0.75

Table 6 The prediction performance for worm genome

Metrics	Methods			
	GRE	3LS	iNuc-PseKNC	iNuc-PseSTNC
Acc	0.8798	0.8786	0.8690	0.8862
S_n	0.8975	0.8654	0.9030	0.9162
S_p	0.8623	0.8921	0.8355	0.8666
Mcct	0.7602	0.7576	0.74	0.77

Table 7 The prediction performance for fly genome

Metrics	Methods			
	GRE	3LS	iNuc-PseKNC	iNuc-PseSTNC
Acc	0.8336	0.8314	0.7997	0.8167
S_n	0.8290	0.8407	0.7831	0.7976
S_p	0.8382	0.8274	0.8165	0.8361
Mcct	0.6672	0.6682	0.60	0.63

was high. Therefore, our proposed method was an effective method in nucleosome positioning.

Table 8 The prediction performance for yeast genome

Metrics	Methods			
	GRE	TNS	iNuc-PhysChem	DNA energy
Acc	1	1	0.967	0.981
S_n	1	1	0.972	0.982
S_p	1	1	0.943	0.980
Mcct	1	1	0.936	0.963

3.4 Prediction results for yeast genomes

Based on tenfold cross-validation, prediction results of yeast genome was obtained (Acc = 1, S_n = 1, S_p = 1 and Mcc = 1). As shown in Table 8, results obtained by GRE model were same as results obtained by TNS model (Awazu 2017), higher than those obtained by iNuc-PhysChem (Chen et al. 2012) (Acc = 0.967), DNA energy (Chen et al. 2016) (Acc = 0.981). Besides, prediction accuracy of our model can reach 1 when parameter k was equal to 60, 70, 80, 90 and 100. It was shown that the distribution of core DNA and linker DNA obtained by GRE model was consistent with the real distribution of core DNA and linker DNA. Compared with those methods provided by Chen, our proposed method can find the key factors influencing nucleosome positioning with lower time complexity. Although the method proposed by Awazu can find key factors influencing nucleosome positioning, its time complexity was high, and our proposed method was an effective method in nucleosome positioning.

3.5 Analysis of factors affecting nucleosome positioning

Using GRE model, core DNA of human, worm, fly and yeast genomes can be recognized in high accuracy. Further-

more, the crucial factors of nucleosome positioning have been found.

Parameter k can reflect the distribution of core DNA and linker DNA to some extent. Besides, compared to the distribution of core DNA and linker DNA obtained by non-optimal parameter values, the distribution obtained by the optimal parameter values was more similar to the real distribution of core DNA and linker DNA in the entire genome. It indicated that parameter k played important roles in nucleosome positioning. Besides, the optimal parameter values were different in human, worm, fly and yeast genomes. It indicated that features of nucleosome positioning were species dependent.

Furthermore, GRE models of four species contained common feature vectors (positive four-nucleotide, positive five-nucleotide, negative five-nucleotide, positive six-nucleotide and negative six-nucleotide sequences). Removing those common feature vectors from ten-dimensional feature vectors, Acc was 0.7881, 0.8400 and 0.7663 for human, worm and fly genomes, respectively. Similarly, Acc was 0.9994, 0.9994, 0.9992, 0.9994 and 0.9994 in yeast genome when parameter k equaled 60, 70, 80, 90 and 100, respectively. It showed that Acc obtained by the common feature vectors was higher than those obtained by those feature vectors which were obtained by removing common feature vectors from all feature vectors. Therefore, these common feature vectors were key factors affecting nucleosome positioning.

Although the position of nucleosomes of four species and crucial factors of nucleosome positioning can be determined based on our proposed model, some additional factors, such as the flexibility of DNA fragments, can also influence nucleosome positioning (Sangaiah et al. 2017; Lu et al. 2017; Zhang et al. 2018a, b). Therefore, in the future, we will combine the information of DNA sequences with their physicochemical properties in nucleosome positioning.

From simulation results, we can get the following trends in nucleosome positioning.

- Nucleosome positioning was species dependent.
- Some common factors, including five-nucleotide and six-nucleotide, were important factors in nucleosome positioning of different species.
- The values of parameter k with the highest prediction Acc were different in different species. It indicated that some unique factors can also affect the position of nucleosomes in different species.

3.6 Analysis of the significance of the parameter k

The parameter k reflected the weight of two different distributions and had important biological significance. For searching for the significance of parameter, prediction results obtained by RE and GRE methods were compared. Besides, the optimal k value of the model was studied when a genetic mutation occurred. Details were as follows.

Table 9 Comparison results obtained by GRE and RE methods in the human genome

Method	Metrics			
	Acc	S_n	S_p	Mcc
RE	0.8690	0.9094	0.8291	0.7406
GRE	0.8778	0.9107	0.8452	0.7573

Table 10 Comparison results obtained by GRE and RE methods in the worm genome

Method	Metrics			
	Acc	S_n	S_p	Mcc
RE	0.8433	0.8781	0.8090	0.6885
GRE	0.8798	0.8975	0.8623	0.7602

Table 11 Comparison results obtained by GRE and RE methods in the fly genome

Method	Metrics			
	Acc	S_n	S_p	Mcc
RE	0.7998	0.7686	0.8316	0.6012
GRE	0.8336	0.8290	0.8382	0.6672

Table 12 Comparison results obtained by GRE and RE methods in the yeast genome

Method	Metrics			
	Acc	S_n	S_p	Mcc
RE	0.9978	0.9979	0.9977	0.9955
GRE	1	1	1	1

In order to find out the importance of weight distribution, prediction results obtained by GRE and RE methods were compared in human, worm, fly and yeast genomes, respectively. The parameter k of GRE method was based on the optimal k value. The comparison results are shown in Tables 9, 10, 11 and 12.

As shown in Tables 9, 10, 11 and 12, we can draw a conclusion that the predictive performance of GRE method was better than that of the RE method. Therefore, in terms of prediction accuracy, the adjustment of the weights of the two distributions was very important. The parameter k played an important role in weight adjustment.

Furthermore, the optimal parameters values in the GRE model were studied when a gene mutation occurred at a certain site of the DNA sequences. In order to ensure that the length of the DNA sequences was unchanged, the gene mutation only considered the replacement of the base pair. It was

found that the optimal parameters values in the GRE model were stable. Therefore, the parameter k was conservative in the same species.

4 Conclusion

In this paper, based on relative entropy, a novel nucleosome positioning method was proposed. Using this method, core DNA of human, worm, fly and yeast was recognized by their sequences. In order to evaluate the quality of model, different nucleosome positioning methods were compared with same benchmark datasets. Experimental results showed that our proposed model was an effective nucleosome positioning method. Besides, five-nucleotide and six-nucleotide sequences were considered as crucial factor in nucleosome positioning. Because some additional factors, such as the flexibility of DNA fragments, can also influence nucleosome positioning. Therefore, in the future, we will combine the information of DNA sequences with their physicochemical properties in nucleosome positioning. Besides, we will apply the generalized relative entropy to comparison of text similarity, the allocation of weight indicators in multi-index evaluation systems and pattern recognition.

Acknowledgements This research is funded by National Natural Science Foundation of China project with Grant No. 61502254, Program for Yong Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region with Grant No. NJYT-18-B10, and Open Funds of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education with Grant No. 93K172018K07.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mandoiu I, Balfe P, Zelikovsky A (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinform 12(Suppl6):S1. https://doi.org/10.1186/1471-2105-12-S6-S1

Awazu A (2017) Prediction of nucleosome positioning by the incorporation of frequencies and distributions of three different nucleotide segment lengths into a general pseudo k-tuple nucleotide composition. Bioinformatics 33(1):42–48. https://doi.org/10.1093/bioinformatics/btw562

Beigi S, Gohari A (2014) Quantum achievability proof via collision relative entropy. IEEE Trans Inf Theory 60(12):7980–7986. https://doi.org/10.1109/TIT.2014.2361632

Benson G (2002) A new distance measure for comparing sequence profiles based on path lengths along an entropy surface. Bioinformatics 18(suppl_2):S44–S53. https://doi.org/10.1093/bioinformatics/18.suppl_2.s44

Berbenetz NM, Nislow C, Brown GW (2010) Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. PLoS Genet. https://doi.org/10.1371/journal.pgen.1001092

Bhasin M, Raghava G (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucl Acids Res 32(suppl_2):W414–W419. https://doi.org/10.1093/nar/gkh350

Chen H, Zhou L (2012) A relative entropy approach to group decision making with interval reciprocal relations based on COWA operator. Group Decis Negot 21(4):585–599. https://doi.org/10.1007/s10726-011-9228-8

Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS ONE 7(10):e47843. https://doi.org/10.1371/journal.pone.0047843

Chen W, Lin H, Chou KC (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst 11(10):2620–2634. https://doi.org/10.1039/C5MB00155B

Chen W, Feng P, Ding H, Lin H, Chou KC (2016) Using deformation energy to analyze nucleosome positioning in genomes. Genomics 107(2–3):69–75. https://doi.org/10.1016/j.ygeno.2015.12.005

Fabris F, Doherty A, Palmer D, de Magalhaes JP, Freitas AA (2018) A new approach for interpreting random forest models and its application to the biology of ageing. Bioinformatics 34(14):2449–2456. https://doi.org/10.1093/bioinformatics/bty087

Flores O, Orozco M (2011) nucleR: a package for nonparametric nucleosome positioning. Bioinformatics 27(15):2149–2150. https://doi.org/10.1093/bioinformatics/btr345

Freeman GS, Lequieu JP, Hinckley DM, de Pablo J (2014) DNA shape dominates sequence affinity in nucleosome formation. Phys Rev Lett 113(16):168101. https://doi.org/10.1103/PhysRevLett.113.168101

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gibb S, Strimmer K (2015) Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis. Bioinformatics 31(19):3156–3162. https://doi.org/10.1093/bioinformatics/btv334

Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30(11):1522–1529. https://doi.org/10.1093/bioinformatics/btu083

Ide H, Umezawa M, Ohwada H (2016) Function prediction of disease-related long intergenic non-coding rna using random forest. In: Proceedings of the 7th international conference on computational systems-biology and bioinformatics. https://doi.org/10.1145/3029375.3029384

Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, Trifonov EN (1996) Nucleosome dna sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol 262(2):129–139. https://doi.org/10.1006/jmbi.1996.0503

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006) Nucleosome positions predicted through comparative genomics. Nat Genet 38(10):1210–1215. https://doi.org/10.1038/ng1878

Ismail H, Saigo H, Dukka K (2017) RF-NR: random forest based approach for improved classification of nuclear receptors. IEEE/ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/tcbb.2017.2773063

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al (2009) The

DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458:362–366. https://doi.org/10.1038/nature07667
Karlekar NP, Gomathi N (2018) OW-SVM: ontology and whale optimization-based support vector machine for privacy-preserved medical data classification in cloud. Int J Commun Syst. https://doi.org/10.1002/dac.3700
Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39:1235–1244. https://doi.org/10.1038/ng2117
Lin W, Ji D, Lu Y (2017) Disorder recognition in clinical texts using multi-label structured SVM. BMC Bioinform 18:75. https://doi.org/10.1186/s12859-017-1476-4
Liu H, Duan X, Yu S, Sun X (2011) Analysis of nucleosome positioning determined by DNA helix curvature in the human genome. BMC Genomics 12:72. https://doi.org/10.1186/1471-2164-12-72
Lu M, Liu S, Kumarsangaiah A (2017) Nucleosome positioning with fractal entropy increment of diversity in telemedicine. IEEE Access 6:33451–33459. https://doi.org/10.1109/ACCESS.2017.2779850
Magliery TJ, Regan L (2005) Sequence variation in ligand binding sites in proteins. BMC Bioinform 6:240. https://doi.org/10.1186/1471-2105-6-240
Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. https://doi.org/10.1101/gr.078261.108
Meng Z, Shen H, Huang H (2018) Search result diversification on attributed networks via nonnegative matrix factorization. Inf Process Manag 54(6):1271–1291. https://doi.org/10.1016/j.ipm.2018.05.005
Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z (2007) Nucleosome positioning signals in genomic DNA. Genome Res. https://doi.org/10.1101/gr.6101007
Petralia F, Wang P, Yang J, Tu Z (2015) Integrative random forest for gene regulatory network inference. Bioinformatics 31(12):i197–i205. https://doi.org/10.1093/bioinformatics/btv268
Polishko A, Ponts N, Le Roch KG, Lonardi S (2012) Normal: accurate nucleosome positioning using a modified gaussian mixture model. Bioinformatics 28(12):i242–i249. https://doi.org/10.1093/bioinformatics/bts206
Rahman R, Otridge J, Pal R (2017) Integratedmrf: random forest-based framework for integrating prediction from different data types. Bioinformatics 33(9):1407–1410. https://doi.org/10.1093/bioinformatics/btw765
Sangaiah AK, Samuel OW, Li X (2017) Towards an efficient risk assessment in software projects—fuzzy reinforcement paradigm. Comput Electr Eng. https://doi.org/10.1016/j.compeleceng.2017.07.022
Sarosi G, Ugajin T (2016) Relative entropy of excited states in two dimensional conformal field theories. J High Energy Phys 2016:114. https://doi.org/10.1007/JHEP07(2016)114
Satchwell SC, Drew HR, Travers AA (1986) Sequence periodicities in chicken nucleosome core DNA. J Mol Biol 191(4):659–675. https://doi.org/10.1016/0022-2836(86)90452-3
Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132(5):887–898. https://doi.org/10.1016/j.cell.2008.02.022
Segal E, Widom J (2009) Poly (DA: DT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol 19(1):65–71. https://doi.org/10.1016/j.sbi.2009.01.004
Shao LH, Li YM, Luo Y, Xi ZJ (2017) Quantum coherence quantifiers based on Renyi α-relative entropy. Commun Theor Phys 67(6):631–636. https://doi.org/10.1088/0253-6102/67/6/631
Sinoquet C (2018) A method combining a random forest-based technique with the modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies. BMC Bioinform 19:106. https://doi.org/10.1186/s12859-018-2054-0
Struhl K, Segal E (2013) Determinants of nucleosome positioning. Nat Struct Mol Biol 20:267–273. https://doi.org/10.1038/nsmb.2506
Taherzadeh G, Zhou Y, Liew AWC, Yang Y (2017) Structure-based prediction of protein-peptide binding regions using random forest. Bioinformatics 34(3):477–484. https://doi.org/10.1093/bioinformatics/btx614
Tahir M, Hayat M (2016) iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of saac and chou's pseaac. Mol BioSyst 12(8):2587–2593. https://doi.org/10.1039/C6MB00221H
Tolstorukov MY, Choudhary V, Olson WK, Zhurkin VB, Park PJ (2008) nuScore: a web-interface for nucleosome positioning predictions. Bioinformatics 24(12):1456–1458. https://doi.org/10.1093/bioinformatics/btn212
Vacic V, Uversky VN, Dunker AK, Lonardi S (2007) Composition profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinform 8:211. https://doi.org/10.1186/1471-2105-8-211
Vernikos GS, Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the salmonella pathogenicity islands. Bioinformatics 22(18):2196–2203. https://doi.org/10.1093/bioinformatics/btl369
Wan S, Mak MW, Kung SY (2013) GOASVM: a subcellular location predictor by incorporating term frequency gene ontology into the general form of Chou's pseudo-amino acid composition. J Theor Biol 323:40–48. https://doi.org/10.1016/j.jtbi.2013.01.012
Wang K, Samudrala R (2006) Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinform 7:385. https://doi.org/10.1186/1471-2105-7-385
Woo S, Zhang X, Sauteraud R, Robert F, Gottardo R (2013) PING 2.0: an R/Bioconductor package for nucleosome positioning using next-generation sequencing data. Bioinformatics 29(16):2049–2050. https://doi.org/10.1093/bioinformatics/btt348
Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP (2010) Predicting nucleosome positioning using a duration Hidden Markov model. BMC Bioinform 11:346. https://doi.org/10.1186/1471-2105-11-346
Yasuda T, Sugasawa K, Shimizu Y, Iwai S, Shiomi T, Hanaoka F (2005) Nucleosomal structure of undamaged DNA regions suppresses the non-specific DNA binding of the XPC complex. DNA Repair 4(3):389–395. https://doi.org/10.1016/j.dnarep.2004.10.008
Yudong Z, Shuihua W, Ping S, Preetha P (2015) Pathological brain detection based on wavelet entropy and Hu moment invariants. Bio-Med Mater Eng 26(s1):S1283–S1290. https://doi.org/10.3233/BME-151426
Zhang YD, Wu LN (2008) Pattern recognition via PCNN and Tsallis entropy. Sensors 8(11):7518–7529. https://doi.org/10.3390/s8117518
Zhang Y, Wu L (2011) Optimal multi-level thresholding based on maximum Tsallis entropy via an artificial bee colony approach. Entropy 13(4):841–859. https://doi.org/10.3390/e13040841
Zhang Y, Gao X, Katayama S (2015) Weld appearance prediction with BP neural network improved by genetic algorithm during disk laser welding. J Manuf Syst 34:53–59. https://doi.org/10.1016/j.jmsy.2014.10.005
Zhang J, Hadj-Moussa H, Storey KB (2016) Current progress of high-throughput microRNA differential expression analysis and random forest gene selection for model and non-model systems:

an R implementation. J Integr Bioinformatics 13(5):35–46. https://doi.org/10.1515/jib-2016-306

Zhang C, Li D, Sangaiah A (2017) Merger and acquisition target selection based on interval neutrosophic multigranulation rough sets over two universes. Symmetry 9(7):126. https://doi.org/10.3390/sym9070126

Zhang J, Peng W, Wang L (2018a) LeNup: learning nucleosome positioning from DNA sequences with improved convolutional neural networks. Bioinformatics 34(10):1705–1712. https://doi.org/10.1093/bioinformatics/bty003/4796955

Zhang C, Li D, Broumi S (2018b) Medical diagnosis based on single-valued neutrosophic probabilistic rough multisets over two universes. Symmetry 10(6):213. https://doi.org/10.3390/sym10060213

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.