This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779850, IEEE Access

Nucleosome Positioning with Fractal Entropy Increment of Diversity in Telemedicine

Mengye Lu, Shuai Liu, Arun Kumarsangaiah, Yunpeng Zhou, Zheng Pan, Yongchun Zuo

ABSTRACT— Recently, telemedicine solutions have become a new trend in remote medical treatment. Many diseases are originated from abnormal variation of biological processes, especially in nucleosome positioning. Thus, effective prediction of nucleosome positioning becomes a hotspot in research of telemedicine. In this paper, a novel method is provided to compare varies of sequences firstly. This method, which called fractal entropy increment of diversity (FEID) is based on information entropy and increment of diversity. Then, a novel nucleosome positioning method is provided by using FEID into the data set of diversiform DNA sequences of human, worm, fly and yeast. Moreover, experimental results show that FEID is an effective nucleosome positioning method by compared with other methods on several benchmark datasets. Finally, the most important nucleotide sequence in nucleosome positioning is provided based on calculated contribution rates of nucleotide sequences.

INDEX TERMS—Telemedicine, information entropy, fractal entropy increment of diversity, fractal, nucleosome positioning.

Ⅰ.INTRODUCTION

Telemedicine is a new medical technology which can provide remote diagnosis, treatment and consultation service to area with poor medical condition. Telemedicine becomes a research hotspot because it helps people who live in remote areas. However, many diseases are related to variation of biological processes which can't be directly observed in telemedicine. Recently, with the development of computer technology, communications technology and medical technology, some information involved data, voice and image can transfer in a long distance. Which contributes to the implementation of telemedicine. In order to obtain more accurate diagnostic results, study the cause of the disease from the gene level is necessary. Besides, many diseases are associated with abnormal reactions in biological processes. In this case, long-distanced nucleosome positioning is valued to study because it affects various bioprocesses [1-2]. Furthermore, analysis of nucleosome positioning is significant for providing an in-depth understanding of biological processes in diseases, such as binding of transcription factors and transcriptional regulation mechanism [3, 4].

Nucleosome is the basic units of eukaryotic chromatin, which is constructed by a histone octamer that wrapped tightly by a DNA sequence with 147 base pair (bp) (Figure 1). Nucleosome positioning provides important information for remote diagnosis, for example, a gene based diagnosis can be quickly determined if genetic variations can be found by remote nucleosome positioning. Hence it becomes a critical issue in telemedicine in recent years.

This work is supported by Programs of National Natural Science Foundation of China (No: 61502254) and The Postgraduate Scientific Research Innovation Foundation of Inner Mongolia (Grant No. 10000-16010109-58)

M. Lu (lmy@mail.imu.edu.cn), S. Liu (cs_liushuai@imu.edu.cn), Y. Zhou (zhouyunpeng1990@126.com) and Z. Pan (cs_pz@imu.edu.cn) are with College of Computer Science, Inner Mongolia University, Hohhot, China and Inner Mongolia Key Laboratory of Social Computing and Data Processing, Hohhot, China.

Arun Kumar Sangaiah (arunkumarsangaiah@gmail.com) is with School of Computing Science and Engineering, VIT University, Vellore, India.

Y. Zuo (yongchunzuo@163.com) is with The State key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Inner Mongolia University, Hohhot, 010070, China

Correspondence will be addressed to Shuai Liu (cs_liushuai@imu.edu.cn)

Recently, studies of theoretical prediction model and experiment of nucleosome positioning have been developed based on development of genetic maps in human, mouse, chicken and yeast genomes [5-7]. Today, many prediction theories of nucleosome positioning have been presented, such as nucleosome-DNA interaction model, N-Score combined with mathematical regression model, curvature spectrum, term Hidden Markov Model and relative fragment frequency index [8-13].

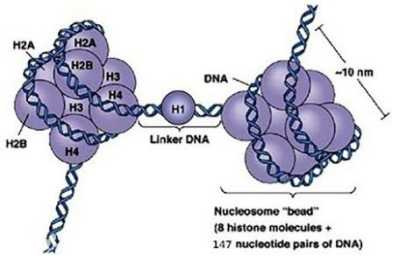


Fig. 1. Nucleosome is constructed by a histone octamer that wrapped tightly by a DNA sequence with 147 base pair (bp). And two nucleosomes are connected by linker DNA

Earlier study by Satchwell et al. found that GG di-nucleotide sequence appeared about 10-bp periodicity in DNA sequences, which was used to bend around histone octamer [3]. Afterwards, various nucleosome positioning methods had been proposed based on DNA sequence scores. Xing et al. used functions related position in nucleosome positioning [11]. Polishko et al. proposed an improved Gaussian model for nucleosome positioning [9]. Xi et al. used Hidden Markov Model in nucleosome positioning and obtained nucleosome occupancy maps by used software "NuPoP" [10]. Besides, based on a probabilistic graphical model, Yassour et al. proposed a new model in nucleosome positioning [12]. Based on the NPS algorithm, iNPS algorithm was used in nucleosome positioning [8, 13]. Furthermore, several nucleosome positioning models were proposed according to the information of di-nucleotide sequences frequencies [5, 6, 14]. Recently, many software and R packages were applied to predict the position of nucleosome [15-17].

Increment of diversity was also an important measure of biological sequences, which was widely utilized to measure similarity of two diversity sources. Earlier, Li et al. predicted the structural class of protein based on increment of diversity [18]. Later, Chen et al. presented a SVM-based method to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779850, IEEE Access

predict subcellular localization according to combined increment of diversity [19]. Increment of diversity and quadratic discriminant analysis (IDQD) was adopted by Yun et al. to predict subcellular localization [20]. Wang et al. constructed immune classifier using increment of diversity [21]. Besides, Yang et al. predicted presynaptic and postsynaptic neurotoxins based on increment of diversity [22]. Zuo et al. used the subsequence increment to predict the plant pol-II promoter [23]. They also used minimum increment of diversity to predict protein amino acids of plasmodium [24]. Afterwards, DNA sequence and protein pattern recognition were studied by increment of diversity and quadratic discriminant analysis [25]. Based on increment of diversity, Chen et al. identified mitochondrial protein of malaria parasites and developed an effective treatment for reducing morbidity of malaria [26]. Recently, Feng et al. have identified antimicrobial peptides by increment of diversity [27]. Zhang et al. trained a SVM detector to do pathological brain detection [28]. And nucleosome positioning, which is important to understand biological processes, has also been studied by increment of diversity and quadratic discriminant analysis in [14]. Furthermore, Zhao et al. calculated scores of different regions in genome and analyzed the nucleosome occupancy [14, 29].

Studies of nucleosome positioning have shown that position of nucleosome is related to genome sequences. Indeed, some sequences of the core DNA appear periodically, which is favorable for the DNA fragments to bend around the histone octamer [3, 30, 31]. Furthermore, a sequence with high affinity to histone is beneficial to the formation of compact structure of nucleosome. However, a DNA sequence with low affinity to histone is helpful when transcription factor approaches target sequence. Therefore, core-DNA can be truly recognized according to the distribution of genome and positions of nucleosome can be determined by DNA sequences.

In this paper, based on the increment of diversity, a new metric "fractal entropy increment of diversity (FEID) [32-34]" is used to find a new nucleosome positioning method by DNA sequence because fractal is a measure of self-similarity and fractal dimension is an important characteristic of objects with complex structure. Thus, when consider that iterative process of genomic structure is related to fractal, we first extract the fractal dimension of DNA sequences as an important feature of nucleosome positioning. Then, with similar hypothesize that gene sequences impact nucleosome positioning [3, 30, 31], we put these characteristics into back propagation neural network (BP neural network) to predict position of nucleosome on human, worm, fly and yeast genomes datasets [35, 36]. Meantime, jackknife test and 10-fold cross-validation are used to evaluate the effectiveness of our model. Finally, analysis of contribution rates of nucleotide sequences is provided to find key factors that influence nucleosome positioning.

Rest of the paper is organized as follows. Section 2 presents materials and methods of this paper, including sources of benchmark datasets, fractal entropy increment of diversity, FEID-BP model and evaluation metrics of results; Section 3 shows the experiment results and its additional analysis; Proposed method and future plans are discussed in Section 4; Conclusion is provided in Section 5.

Ⅱ.MATERIALS AND METHODS

A. DATASET

In this paper, core-DNA distributions of different species are analyzed with a combined dataset, including nucleosome information in both human, worm, fly and yeast genomes. The dataset of human comes from Guo et al. [6], and compared nucleosome positioning data is from Schones et al. [37] (http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx). The data of human genome used in this paper is "hg 18" version of UCSC human genome database (http://hgdownload.cse.ucsc.edu/). Partial data extracted from chromosome 20 of human genome is used as experimental nucleosome positioning dataset of human as follows [38].

First, each DNA fragment is assigned with a given nucleosome formation score. Then, the sequences with highest scores are chosen as core DNAs, while those with the lowest scores are chosen as linker DNAs. Finally, to reduce redundant data in current dataset, CD-HIT software is applied to remove redundancy with threshold 80% [39]. The obtained benchmark dataset contains 2273 core DNA sequences and 2300 linker DNA sequences with same length 147 bp. The dataset is same to dataset in supplementary data of Ref.6.

Similarly, the dataset of worm and fly are from Guo et al. [6]. Entire genome data are available from the UCSC genome database (http://hgdownload.cse.ucsc.edu/). Compared nucleosome positioning data of worm is also from UCSC genome database and compared nucleosome positioning data of fly is from Mavrich et al. (http://atlas.bx.psu.edu/) [40, 41]. In UCSC database, WS170/ce4 version is chosen as entire worm genome data and BDGP R5 version is chosen as entire fly genome data in this paper. Then, by same methodology to construct dataset of human genome, experimental nucleosome positioning datasets of worm and fly are finally obtained.

Dataset of worm contains 2567 core DNA sequences and 2608 linker DNA sequences with same length 147 bp. The dataset is same to dataset in supplementary data of Ref.6. Dataset of fly contains 2900 core DNA sequences and 2850 linker DNA sequences with same length 147 bp. The dataset is same to dataset in supplementary data of Ref.6.

Dataset of the yeast genome is constructed by Chen et al. [42] and compared nucleosome positioning data is from Lee et al. [7]. With entire genome data (http://www.yeastgenome.org/) and same chosen strategy, benchmark dataset is obtained with 1880 core DNA sequences and 1740 linker DNA sequences with same length 150 bp. The dataset is same to dataset in supporting information of Ref.43.

B. FRACTAL ENTROPY INCREMENT OF DIVERSITY

1) INCREMENT OF DIVERSITY

Measure of Diversity (MD) is applied to describe interaction of factors in high dimensional space $S = \{m_1, m_2, \dots m_s\}$, which is composed of s different dimensions. Let $X \in S$; x_i denotes frequency of i^{th} dimension of X in simple base; when assuming $(\forall x) 0 \cdot \log_b \frac{0}{x} = 0$ is tautology, measure of diversity MD(X) of $X(x_1, x_2, \dots, x_s)$ is defined as Eq.1, where $N_X = \sum_{i=1}^{s} x_i$ is frequencies' amount of every x_i in X; b is the given base of logarithm.

$$MD(X) = N_X \log_b N_X - \sum_{i=1}^{s} x_i \log_b x_i \quad (1)$$

Similarly, when we have another diversity source $Y(y_1, y_2, \dots, y_s) \in S$, MD(Y) can be defined as Eq.2, where y_i denotes frequency of i^{th} component in Y and $N_Y = \sum_{i=1}^{t} y_i$.

$$MD(Y) = N_Y \log_b N_Y - \sum_{i=1}^{s} y_i \log_b y_i \quad (2)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779850, IEEE Access

Then, MD of X+Y is defined as Eq.3.

$$MD(X+Y) = (N_X + N_Y)\log_b(N_X + N_Y) - \sum_{i=1}^{s}(x_i + y_i)\log_b(x_i + y_i) \quad (3)$$

Thus, increment of diversity (ID) of X and Y is defined to measure similarity of two diversity sources X and Y by Eq.4, which is widely used to describe similarity between X and Y. The more similar between X and Y, the less ID(X,Y) is calculated.

$$ID(X,Y) = MD(X+Y) - MD(X) - MD(Y) \quad (4)$$

2) ENTROPY INCREMENT OF DIVERSITY

In this paper, based on increment of diversity and information entropy in information space, Entropy Increment of Diversity (EID) is provided.

We have Eq.5 to describe diversity of sequence X, where H(X) denotes information entropy of sequence X. H(X) reflects chaos of DNA sequences. The larger the value is, the more chaotic the DNA sequences are. The meaning of Nx is the same as the Nx in Eq. 1, and the value is equal to the length of DNA sequences minus 1.

$$\mathrm{MD(X)} = \mathrm{N_X H(X)} \quad (5)$$

Because core DNA sequences and linker DNA sequences have different sequence preferences. Besides, different features of these two kinds of sequences become apparent if we splice those DNA sequences with same sequence preferences. Thus, it is necessary to consider length of DNA sequence when we measure the diversity of DNA sequences, which means that MD(X) reflects the sequence preference of DNA sequence.

Assuming that there is another sequence $Y(y_1, y_2, \dots, y_s)$, which is composed of same s different components, we have Eq.6 to record MD(Y) like Eq.5.

$$\mathrm{MD(Y)} = \mathrm{N_Y H(Y)} \quad (6)$$

When we have MD(X+Y) in Eq.7, EID(X, Y) can be defined in Eq.8 where k=0 if X=Y, k=1 if X≠Y, $\mathrm{N} = \min(N_x, N_y)$.

$$\mathrm{MD(X+Y)} = \mathrm{N_{X+Y} H(X+Y)} \quad (7)$$

$$\mathrm{EID(X,Y)} = \mathrm{MD(X+Y) - MD(X) - MD(Y) + 2kNlogs} \quad (8)$$

Each DNA sequence has a certain sequence preference. Type of the DNA sequences can be determined by calculated their sequence preference. Meanwhile, in case that results of EID are negative, a constant was added in Eq.8.

3) FRACTAL DIMENSION

In this paper, fractal dimension is another method used to measure similarity. Box-counting method is applied to calculate fractal dimension of DNA sequences.

Supposing A as a nonempty bound subset of space R_n, a number of boxes are applied to cover A. For all r>0, $N_r(A)$, which denotes minimum boxes' number, is given by Eq.9 where $\mathrm{d} = \lim_{r\to 0}\frac{\log N_r(A)}{\log(\frac{1}{r})}$ means box-counting dimension.

$$\mathrm{N_r}(A) \propto \frac{1}{r^d} \quad (9)$$

4) PREDICTING MODEL: FEID-BP MODEL

In this paper, with calculated FEID of various types of DNA fragments in both core DNA and linker DNA, a recognition method of nucleosome positioning is provided by used BP neural network which is constructed as follows.

Step 1. Combine all core DNA sequences as one long sequence, count sequence frequencies of all types of di-nucleotide and calculate its MD as $X_2 = \{x_1, x_2, \dots x_{16}\}$.

Step 2. Combine all linker DNA sequences as one long sequence, count sequence frequencies of all types of di-nucleotide and calculate its MD as $Y_2 = \{y_1, y_2, \dots y_{16}\}$.

Step 3. Count sequence frequencies of all types of di-nucleotide in all core DNA sequences and all linker DNA sequences. For each DNA sequences (assuming its serial number is k), calculate its MD $X_k' = \{x_{k,1}', x_{k,2}', \dots x_{k,16}'\}$.

Step 4. Calculate both $s_{k,2}^{+}$ =EID(X_k',X_2) and $s_{k,2}^{-}$=EID(X_k',Y_2) by Eq.8 to obtain a two dimensions feature vectors for each X'. Calculate the box-counting dimensions $s_{k,d}$ for each core DNA and linker DNA by Eq.9.

Step 5. Similarly, calculate fractal entropy increment of diversity of other types of DNA fragments, including tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide fragments. For each DNA sequence with serial number k, we have a vector S_k with 11 dimensions in Eq.10.

$$S_k = \left\{s_{k,2}^{+}, s_{k,2}^{-}, s_{k,3}^{+}, s_{k,3}^{-}, s_{k,4}^{+}, s_{k,4}^{-}, s_{k,5}^{+}, s_{k,5}^{-}, s_{k,6}^{+}, s_{k,6}^{-}, s_{k,d}\right\} \quad (10)$$

Step 6. Use S_k as feature vector of k^{th} DNA sequence k, recognized all S_k with BP neural network.

5) EVALUATIONS OF THE QUALITY OF PREDICTION

Independent data test, k-fold cross-validation and jackknife test are often used to evaluate model's quality. However, jackknife test is deemed to have a better effect than other methods [43]. Thus, jackknife test is widely used by many scholars to evaluate the quality of a model [5, 6, 44-49].

In this paper, four factors TP, FP, FN and TN are defined as true positive, false positive, false negative and true negative, respectively. Detailedly, TP denotes the number that core DNA sequences are predicted to core DNA sequences; TN denotes the number that linker DNA sequences are predicted to linker DNA sequences; FP denotes the number that linker DNA sequences are predicted to core DNA sequences; FN denotes the number that core DNA sequences are predicted to linker DNA sequences. The following metric S_n, S_p, Acc and Mcc are defined to evaluate performance of our method, where S_n denotes sensitivity, S_p denotes specificity, Acc denotes accuracy, and Mcc denotes Mathew correlation coefficient [5].

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

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

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

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

III. RESULTS

A. CONTRIBUTION RATES OF THE NUCLEOTIDE SEQUENCES

Contribution rate is an important metric to measure benefits' factors in economics. When assume x is input and y is output; quantity of x and y are 0 in the initial state; y is a function of x; contribution ratio cr is defined by Eq.11.

$$\mathrm{cr} = \frac{y(x)}{x} \quad (11)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779850, IEEE Access

Based on Eq.11, contribution rate is defined as Eq.12 in arbitrary initial state. ∇y and ∇x are defined as increment of output and input, respectively.

$$\nabla \mathrm{cr} = \frac{\nabla y}{\nabla x} \tag{12}$$

According to Eq.12, contribution rates of nucleotide sequences can be calculated as follows.

Step 1. Combine all types of DNA fragments as input factors (di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences) and calculate provided model's accuracy acc by using FEID-BP model.

Step 2. Calculate provided model's accuracy by using each type of DNA fragments as the only input factor. Accuracy of di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences are defined as acc_2, acc_3, acc_4, acc_5 and acc_6, respectively.

Step 3. Combine all types of DNA fragments as input factors except di-nucleotide and calculate the model's accuracy acc_2'. Similarly, $acc_3', acc_4', acc_5', acc_6'$ denote accuracy of our model when remove tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences from input factors, respectively.

Step 4. Calculate contribution rates of different DNA fragments by Eq.13.

$$\nabla cr_i = \frac{acc - acc_i'}{acc_i} \tag{13}$$

Here, ∇cr_i is contribution rate of each type of i-nucleotide, and $i = \{2, 3, 4, 5, 6\}$. Detailedly, contribution rates of di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences are respectively defined as $\nabla cr_2, \nabla cr_3, \nabla cr_4, \nabla cr_5, \nabla cr_6$.

Accuracy and contribution rates of all types of DNA fragments are shown in both Figure 2 and Table 1. In Figure 2, red line means accuracy of FEID by used only one type of DNA fragment, green line means accuracy of FEID by used all type of DNA fragment except one type, blue line means accuracy of FEID by used all type of DNA fragment. In order to make comparison with the red lines and green lines, the lines with different colors are placed in same figure. Values of Y axis have no function linkage to the value of X axis in blue lines because blue line means accuracy using all types of DNA fragments as input factors that it has no function linkage to all types of DNA fragment.

In Table 1, we find that contribution rate of six-nucleotide sequence is highest in human, worm and fly datasets. However, in yeast dataset, contribution rate of five-nucleotide sequences is the highest. Though there are many combinations of nucleotide sequences can be used as input factors based on di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences, accuracy will drop if one type of DNA fragments with negative $\nabla \mathrm{cr_i}$ is put into input factors of FEID-BP model. Moreover, every contribution rate of one type of DNA fragments is not high, which means structure of nucleosome is controlled by many types of nucleotide fragments. In this paper, because different DNA fragments interact with each other and contribution rates of input nucleotide fragments is complex, all types of nucleotide fragments are used in our nucleosome positioning model at last. In fact, accuracy of our model with all 5 types of DNA fragments is larger than accuracy without any type of DNA fragment.

TABLE 1: Contribution rates of different DNA fragments in four species

Species	cr_2(%)	cr_3(%)	cr_4(%)	cr_5(%)	Cr_6(%)
Human	0.76	-0.24	-0.34	0.49	**2.25**
Worm	0.27	-0.28	-0.02	1.32	**3.48**
Fly	0.83	5.77	0.73	2.82	**8.88**
Yeast	-0.03	0.11	0.13	**0.19**	0.05

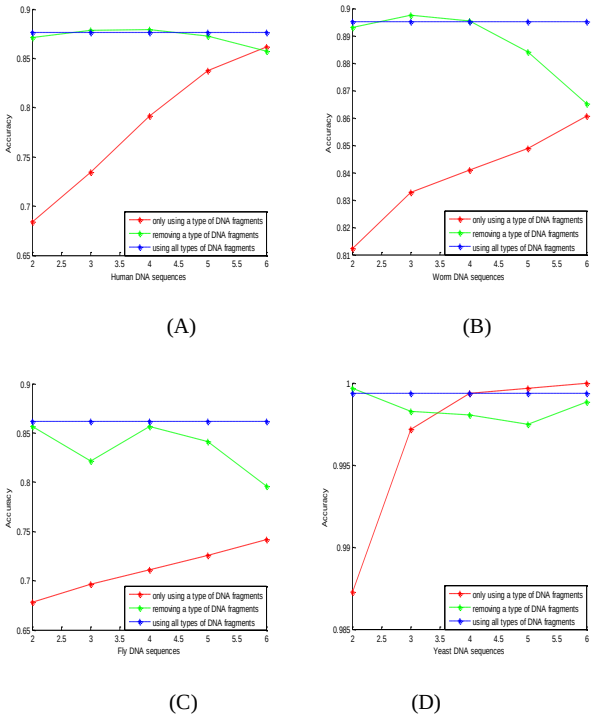


Fig. 2. Accuracy of our FEID-BP with different DNA fragments in human (A), worm (B), fly (C) and yeast (D) genomes. Point at red line denotes accuracy when only use di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences as input factors, respectively. Point at green line denotes accuracy when use all type of DNA fragments except di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences, respectively. Point at blue line denotes accuracy when use all types of DNA fragments (di-nucleotide, tri-nucleotide, four-nucleotide, five-nucleotide and six-nucleotide sequences).

B. PREDICTED RESULTS OF HUMAN,WORM, FLY AND YEAST DATASETS

We use jackknife test to evaluate model's quality with human, worm and fly datasets. Meanwhile, 10-fold cross-validation is used to evaluate model's quality with yeast dataset. Accuracy of FEID is 0.8789, 0.8976, 0.8550 and 0.9994 for human, worm, fly and yeast datasets, respectively. Detailed comparisons are shown in Tables 2-5 and Figure 3. The accuracies of FEID model are higher than other methods with human, worm and fly datasets. Accuracy for yeast dataset reaches 0.9994, higher than those obtained using DNA deformation energy model [50] and iNuc-PhysChem model [51] (Figure 3D).Furthermore, our model has best Mathew correlation coefficient with human, worm and fly datasets which means input factors in this paper is more related to nucleosome positioning. All these results show that FEID-BP model is an effective method in nucleosome positioning.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779850, IEEE Access

TABLE 2. Comparison using jackknife test in human genome

Species	Methods	Acc	Sn	Sp	Mcc
Human	FEID-BP	**0.8789**	**0.9006**	0.8574	**0.7585**
	iNuc-PseKNC [6]	0.8627	0.8786	0.8470	0.73
	iNuc-PseSTNC [49]	0.8760	0.8931	**0.8591**	0.75

TABLE 3. Comparison using jackknife test in worm genome

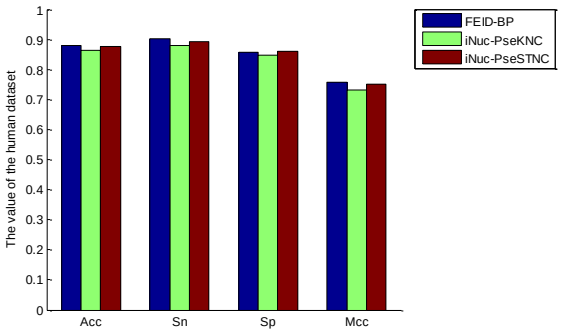
Species	Methods	Acc	Sn	Sp	Mcc
Worm	FEID-BP	**0.8976**	0.9061	0.8892	**0.7953**
	3LS [5]	0.8786	0.8654	**0.8921**	0.7576
	iNuc-PseKNC [6]	0.8690	0.9030	0.8355	0.74
	iNuc-PseSTNC [49]	0.8862	**0.9162**	0.8666	0.77

TABLE 4. Comparison using jackknife test in fly genome

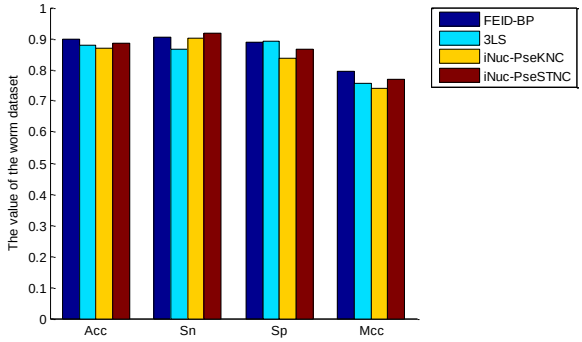
Species	Methods	Acc	Sn	Sp	Mcc
Fly	FEID-BP	**0.8550**	**0.8479**	**0.8621**	**0.7100**
	3LS [5]	0.8341	0.8407	0.8274	0.6682
	iNuc-PseKNC [6]	0.7997	0.7831	0.8165	0.60
	iNuc-PseSTNC [49]	0.8167	0.7976	0.8361	0.63

TABLE 5. Comparison using 10-fold cross-valudation test in yeast genome

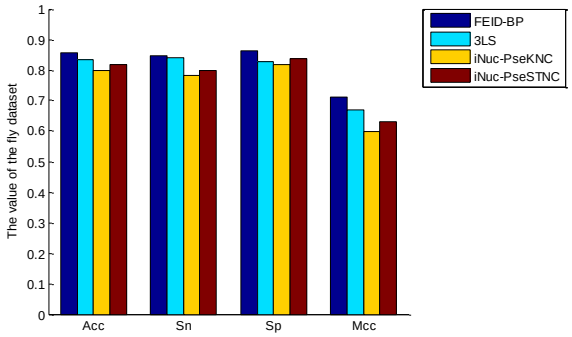
Species	Methods	Acc	Sn	Sp	Mcc
Yeast	FEID-BP	0.9994	**1.0**	0.9989	0.9989
	3LS [5]	**1.0**	**1.0**	**1.0**	**1.0**
	DNA deformation energy [50]	0.981	0.982	0.980	0.963
	iNuc-PhysChem [51]	0.967	0.972	0.943	0.936



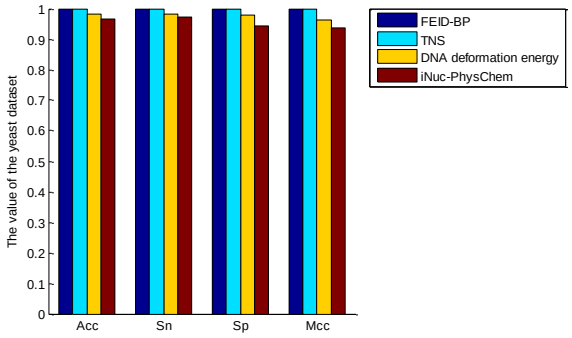
(A)



(B)



(C)



(D)

Fig. 3. Comparison results of various methods in human (A), worm (B), fly (C) and yeast (D) genomes datasets. X axis denotes the metrics to evaluate model's quality, including Acc, Sn, Sp and Mcc, and Y axis deno,tes the values of four metrics in human, worm, fly and yeast genomes, respectively.

Besides, the iNuc-PseKNC model and the iNuc-PhysChem model are based on SVM, those methods may be deficient in finding key factors in nucleosome positioning. While the most important factors in nucleosome positioning can be found based on calculated contribution rates of nucleotide sequences.

Ⅳ. DISCUSSION

From Table 1, we find that six-nucleotide plays more important role in nucleosome positioning than other DNA fragments. We believe it is because of two main reasons.

On the one hand, a six-nucleotide sequence is composed by two molecule tri-nucleotide sequences. We know that a tri-nucleotide sequence can participates in gene expression

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2779850, IEEE Access

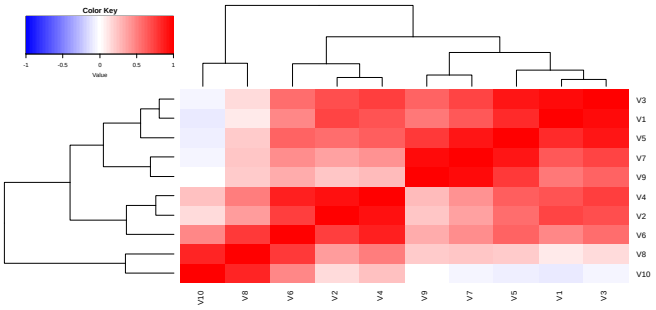
processes because they can construct codons in transcription [52]. In the processes of gene expression, the changes of nucleosome are quite clear. Furthermore, nucleosome has a low occupancy in TSS and CTCF position [8].

On the other hand, six-nucleotide sequences contain more sequences information in nucleosome positioning. Thus, we deduce that characteristics of six-nucleotide sequences are similar to feature of nucleosome [53]. In the future, we will combine the six-nucleotide sequences with six dimensional flexibilities of nucleotide sequences in nucleosome positioning. Besides, based on the distribution map of nucleosome positioning, important information for remote diagnosis will be found.

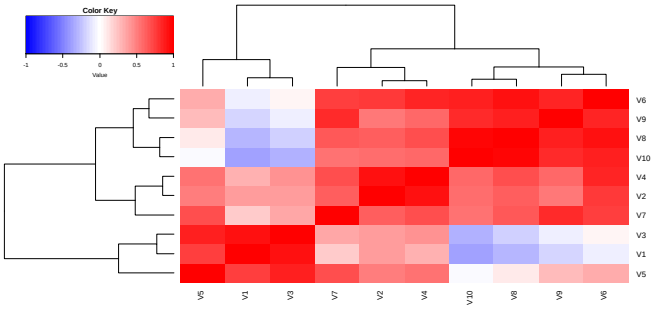
Besides, we analyze relevance of the ten features in order to find more factors related to nucleosome positioning.

Correlation degree can be measured by correlation coefficient. Due to the unknown distribution of ten features, spearman correlation coefficient is used to calculate relevance of ten features.

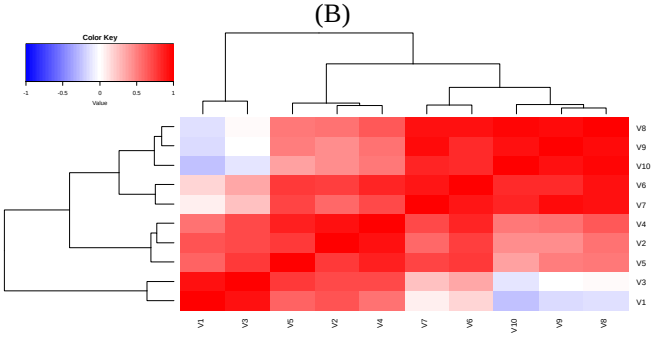
From Figure 4A,B,C,D, we can obtain relevance of ten features in human, worm, fly and yeast genomes, respectively.



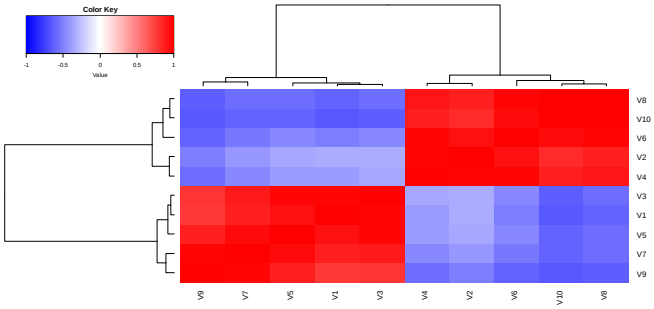
(A)



(B)



(C)



(D)

Fig. 4. Relevance of ten features in human, worm, fly and yeast genomes were shown by color matrix. Different color denoted different correlation coefficient, the relation between color and correlation coefficient was shown in legend. For simplicity, v1~v10 denoted ten features vectors obtained by positive di-nucleotide, negative di-nucleotide, positive tri-nucleotide, negative tri-nucleotide, positive four-nucleotide, negative four-nucleotide, positive five-nucleotide, negative five-nucleotide, positive six-nucleotide and negative six-nucleotide, respectively. Furthermore, those features vectors with high relevance were connected by a line and the relevance became weaker and weaker with the increase of distance with color matrix. (a) Relevance of ten features in human genome; (b) Relevance of ten features in worm genome; (c) Relevance of ten features in fly genome; (d) Relevance of ten features in yeast genome.

As shown in Figure 4, some features vectors are highly relevant. In order to obtain accurate results, meanwhile, reduce calculation time, the following strategy was adopted.

At first, search for those features vectors with high relevance in four species genomes and set relevant threshold. And two features vectors whose correlation coefficient exceeded the threshold were considered as features vectors of highly relevant. Then, use all features vectors except those features vectors with highly relevant as input feature vector of model to recognize core DNA. Finally, 10-fold cross-validation was used to examine the performance of prediction model. The threshold and feature vectors with high velevance in Table 6. The prediction results obtained all feature vectors except those with high relevance were shown in Table 7.

TABLE 6. Feature vectors with high relevance in four species

Species	threshold	feature vectors with high relevance
Human	90%	v1 and v3, v2 and v4, v3 and v5, v5 and v7, v7 and v9.
worm	90%	v1 and v3, v2 and v4, v6 and v8, v8 and v10.
fly	93%	v2 and v4, v7 and v9, v8 and v9, v8 and v10, v9 and v10.
yeast	95%	v1 and v3, v2 and v4, v3 and v5, v4 and v6, v5 and v7, v6 and v8, v7 and v9, v8 and v10.

TABLE 7. Prediction results obtained by feature vectors combinations as input feature vectors of model

Species	new feature vectors	Acc	Sn	Sp	Mcc
Human	v1,v4,v5,v6, v8,v9,v10.	0.8725	0.9039	0.8418	0.7468
Worm	v2,v3,v5,v6, v7,v9,v10.	0.8902	0.9081	0.8733	0.7814
Fly	v1,v3,v4,v5, v6,v7,v8.	0.7970	0.7843	0.8105	0.5949
Yeast	v1,v2,v5,v6, v9,v10	0.9981	0.9990	0.9971	0.9961

As shown in Table 6-7, using new feature vectors with low relevance as input vectors of model, the prediction accuracy of human, worm and yeast was slightly lower than those obtained by 10 feature vectors. Which illustrated that new features of Table 7 were important factors in nucleosome positioning. As for fly genome, the prediction results were not ideal only using new feature vectors as input feature vectors of model. Which showed that different DNA fragments interacted with each other and interaction information was favor to nucleosome positioning.

Ⅴ. CONCLUSION

In this paper, a novel nucleosome positioning method based on both fractal, entropy information and increment of diversity was proposed. Based on this method, core DNAs of human, worm, fly and yeast were recognized by their sequences. In order to evaluate model's quality, different nucleosome positioning methods were compared with same exist benchmark datasets. Experimental results showed that the provided model was an effective nucleosome positioning method. Besides, this paper analyzed importance of all factors which were thought to play roles in nucleosome structure. We found six-nucleotide sequences palyed most important role in nucleosome positioning based on analysis of nucleotide sequences we used.

Because nucleosome positioning has great significance in telemedicine, when combined the information of nucleosome positioning with biology processes, genetic variations in remote area will be found recently. Furthermore, a gene based diagnosis will be quickly determined in future.

However, FEID-BP model was based on sequences information, some factors influencing nucleosome position were ignored. In order to obtain more accurate prediction results, we will combine the sequences information with physicochemical properties of sequences in nucleosome positioning in the future.

Acknowledgement

We want to thank Dr. Ma from Surrey University, UK to help grammar proof of this paper.

Supplementary Materials:

Supplementary data is available at https://academic.oup.com/bioinformatics/article/30/11/1522/283594/iNuc-PseKNC-a-sequence-based-predictor-for

Supporting Information S1 is available at http://dx.doi.org/10.1016/j.ygeno.2015.12.005

REFERENCES

[1] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, et al., "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, p. 362, 2009.

[2] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, et al., "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nat. Genet*, vol. 39, p. 311, 2007.

[3] I. P. Ioshikhes, I. Albert, S. J. Zanton, and B. F. Pugh, "Nucleosome positions predicted through comparative genomics," *Nat. Genet*, vol. 38, p. 1210, 2006.

[4] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, et al., "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning," *Genome Res*, vol. 18, pp. 1051-1063, 2008.

[5] A. Awazu, "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*, vol. 33, pp. 42-48, 2016.

[6] S.-H. Guo, E.-Z. Deng, L.-Q. Xu, H. Ding, H. Lin, W. Chen, et al., "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, pp. 1522-1529, 2014.

[7] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, et al., "A high-resolution atlas of nucleosome occupancy in yeast," *Nat. Genet*, vol. 39, p. 1235, 2007.

[8] W. Chen, Y. Liu, S. Zhu, C. D. Green, G. Wei, and J.-D. J. Han, "Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data," *Nat. Commun*, vol. 5, p. 4909, 2014.

[9] A. Polishko, N. Ponts, K. G. Le Roch, and S. Lonardi, "NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model," *Bioinformatics*, vol. 28, pp. i242-i249, 2012.

[10] L. Xi, Y. Fondufe-Mittendorf, L. Xia, J. Flatow, J. Widom, and J.-P. Wang, "Predicting nucleosome positioning using a duration Hidden Markov Model," *BMC Bioinformatics*, vol. 11, p. 346, 2010.

[11] Y. Xing, X. Zhao, and L. Cai, "Prediction of nucleosome occupancy in Saccharomyces cerevisiae using position-correlation scoring function," *Genomics*, vol. 98, pp. 359-366, 2011.

[12] M. Yassour, T. Kaplan, A. Jaimovich, and N. Friedman, "Nucleosome positioning from tiling microarray data," *Bioinformatics*, vol. 24, pp. i139-i146, 2008.

[13] Y. Zhang, H. Shin, J. S. Song, Y. Lei, and X. S. Liu, "Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq," *BMC Genomics*, vol. 9, p. 537, 2008.

[14] X. Zhao, Z. Pei, J. Liu, S. Qin, and L. Cai, "Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis," *Chromosome Res*, vol. 18, pp. 777-785, 2010.

[15] O. Flores and M. Orozco, "nucleR: a package for non-parametric nucleosome positioning," *Bioinformatics*, vol. 27, pp. 2149-2150, 2011.

[16] M. Y. Tolstorukov, V. Choudhary, W. K. Olson, V. B. Zhurkin, and P. J. Park, "nuScore: a web-interface for nucleosome positioning predictions," *Bioinformatics*, vol. 24, pp. 1456-1458, 2008.

[17] S. Woo, X. Zhang, R. Sauteraud, F. Robert, and R. Gottardo, "PING 2.0: an R/Bioconductor package for nucleosome positioning using next-generation sequencing data," *Bioinformatics*, vol. 29, pp. 2049-2050, 2013.

[18] Q.-Z. Li and Z.-Q. Lu, "The prediction of the structural class of protein: application of the measure of diversity," *J. Theor. Biol*, vol. 213, pp. 493-502, 2001.

[19] Y. Chen, Q. Li, K.-l. YANG, and G.-l. FAN, "Predicting of the subcellular location of apoptosis proteins using the algorithm of the increment of diversity combined with support vector machine," *Acta Biophysica Sinica*, vol. 23, pp. 192-198, 2007.

[20] L. Wang, J. Zhang, G. Gong, and M. Peng, "Application of immune classifier based on increment of diversity in the model species genomes identification," Presented at Proc. International Conference on Intelligent Computation Technology and Automation (ICICTA).

[21] J. Yun, J. Zhao, and L. Jun, "The recognition of subcellular localization of proteins based on their N-terminal amino acid sequence," Presented at International Conference on Bioinformatics & Biomedical Engineering.

[22] L. Yang and Q. Li, "Prediction of presynaptic and postsynaptic neurotoxins by the increment of diversity," *Toxicol. Vitro*, vol. 23, pp. 346-348, 2009.

[23] Y. Zuo and Q. Li, "Predicting plant pol-II promoter based on subsequence increment of overlap content diversity," Presented at Proc. 2nd International Conference on Biomedical Engineering and Informatics.

[24] Y.-C. Zuo and Q.-Z. Li, "Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids," *Amino Acids*, vol. 38, pp. 859-867, 2010.

[25] J. Lu, L. Luo, L. Zhang, W. Chen, and Y. Zhang, "Increment of diversity with quadratic discriminant analysis-an efficient tool for sequence pattern recognition in bioinformatics," *Open Access Bioinf*, vol. 2, pp. 89-96, 2010.

[26] Y.-L. Chen, Q.-Z. Li, and L.-Q. Zhang, "Using increment of diversity to predict mitochondrial proteins of malaria parasite: integrating pseudo-amino acid composition and structural alphabet," *Amino Acids*, vol. 42, pp. 1309-1316, 2012.

[27] P. Feng, Z. Wang, and X. Yu, "Predicting antimicrobial peptides by using increment of diversity with quadratic discriminant analysis method," *IEEE/ACM Trans. Computat. Biol. Bioinformatics*, 2017.

[28] Y.-D. Zhang, Y. Jiang, W. Zhu, S. Lu, and G. Zhao, "Exploring a smart pathological brain detection method on pseudo Zernike moment," *Multimed. Tools Appl*, pp. 1-16, 2017.

[29] W. Chen, L. Luo, and L. Zhang, "The organization of nucleosomes around splice sites," *Nucleic Acids Res*, vol. 38, pp. 2788-2798, 2010.

[30] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E. N. Trifonov, "Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences," *J Mol Biol*, vol. 262, pp. 129-139, 1996.

[31] S. C. Satchwell, H. R. Drew, and A. A. Travers, "Sequence periodicities in chicken nucleosome core DNA," *J Mol Biol*, vol. 191, pp. 659-675, 1986.

[32] S. Liu, X. Cheng, C. Lan, W. Fu, J. Zhou, Q. Li, et al., "Fractal property of generalized M-set with rational number exponent," *Appl. Math. Comput*, vol. 220, pp. 668-675, 2013.

[33] S. LIU, Z. PAN, and X. CHENG, "A NOVEL FAST FRACTAL IMAGE COMPRESSION METHOD BASED ON DISTANCE CLUSTERING IN HIGH DIMENSIONAL SPHERE SURFACE," *Fractals*, p. 1740004, 2017.

[34] S. Liu, Z. Pan, W. Fu, and X. Cheng, "Fractal generation method based on asymptote family of generalized Mandelbrot set and its application," *J. Nonlinear Sci. Appl*, vol. 10, pp. 1148-1161, 2017.

[35] S. Wang, R. V. Rao, P. Chen, Y. Zhang, A. Liu, and L. Wei, "Abnormal breast detection in mammogram images by feed-forward neural network trained by Jaya algorithm," *Fundam. Inform*, vol. 151, pp. 191-211, 2017.

[36] Y.-D. Zhang, Y. Zhang, Y.-D. Lv, X.-X. Hou, F.-Y. Liu, W.-J. Jia, et al., "Alcoholism detection by medical robots based on Hu moment invariants and predator–prey adaptive-inertia chaotic particle swarm optimization," *Comput. Electr. Eng*, 2017.

[37] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, et al., "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, pp. 887-898, 2008.

[38] H. Liu, X. Duan, S. Yu, and X. Sun, "Analysis of nucleosome positioning determined by DNA helix curvature in the human genome," *BMC Genet*, vol. 12, p. 72, 2011.

[39] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, pp. 3150-3152, 2012.

[40] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, et al., "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome Res*, vol. 18, pp. 1073-1083, 2008.

[41] T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, et al., "Nucleosome organization in the Drosophila genome," *Nature*, vol. 453, p. 358, 2008.

[42] W. Chen, H. Lin, and K.-C. Chou, "Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences," *Mol. Biosyst*, vol. 11, pp. 2620-2634, 2015.

[43] K.-C. Chou and H.-B. Shen, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nat. Protoc*, vol. 3, p. 153, 2008.

[44] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou,

"Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS One*, vol. 7, p. e35254, 2012.

[45] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res*, vol. 41, pp. e68-e68, 2013.

[46] Y.-K. Chen and K.-B. Li, "Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition," *J. Theor. Biol*, vol. 318, pp. 1-12, 2013.

[47] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. Biosyst*, vol. 8, pp. 629-641, 2012.

[48] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *J. Theor. Biol*, vol. 263, pp. 203-209, 2010.

[49] M. Tahir and M. Hayat, "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*, vol. 12, pp. 2587-2593, 2016.

[50] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "Using deformation energy to analyze nucleosome positioning in genomes," *Genomics*, vol. 107, pp. 69-75, 2016.

[51] W. Chen, H. Lin, P.-M. Feng, C. Ding, Y.-C. Zuo, and K.-C. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS One*, vol. 7, p. e47843, 2012.

[52] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, "Codon catalog usage is a genome strategy modulated for gene expressivity," *Nucleic Acids Res*, vol. 9, pp. 213-213, 1981.

[53] G.-C. Yuan and J. S. Liu, "Genomic sequence is highly predictive of local nucleosome depletion," *PLOS Comput. Biol*, vol. 4, p. e13, 2008.