Nucleosome Positioning Based on the Sequence Word Composition

Xian-Fu Yi[1,2], Zhi-Song He[3], Kuo-Chen Chou[4] and Xiang-Yin Kong[1,2,*]

[1]*State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, 197 Rui Jin Road II, Shanghai 200025, China;* [2]*Institute of Health Sciences, Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS) and Shanghai Jiao Tong University School of Medicine (SJTUSM), 225 South Chong Qing Road, Shanghai 200025, China;* [3]*Department of Bioinformatics, College of Life Sciences, Zhejiang University, HangZhou, ZheJiang 310058, China;* [4]*Gordon Life Science Institute, San Diego, California 92130, USA*

Abstract: The DNA of all eukaryotic organisms is packaged into nucleosomes (a basic repeating unit of chromatin). A nucleosome consists of histone octamer wrapped by core DNA and linker histone H1 associated with linker DNA. It has profound effects on all DNA-dependent processes by affecting sequence accessibility. Understanding the factors that influence nucleosome positioning has great help to the study of genomic control mechanism. Among many determinants, the inherent DNA sequence has been suggested to have a dominant role in nucleosome positioning *in vivo*. Here, we used the method of minimum redundancy maximum relevance (mRMR) feature selection and the nearest neighbor algorithm (NNA) combined with the incremental feature selection (IFS) method to identify the most important sequence features that either favor or inhibit nucleosome positioning. We analyzed the words of 53,021 nucleosome DNA sequences and 50,299 linker DNA sequences of *Saccharomyces cerevisiae*. 32 important features were abstracted from 5,460 features, and the overall prediction accuracy through jackknife cross-validation test was 76.5%. Our results support that sequence-dependent DNA flexibility plays an important role in positioning nucleosome core particles and that genome sequence facilitates the rapid nucleosome reassembly instead of nucleosome depletion. Besides, our results suggest that there exist some additional features playing a considerable role in discriminating nucleosome forming and inhibiting sequences. These results confirmed that the underlying DNA sequence plays a major role in nucleosome positioning.

Keywords: DNA flexibility, feature selection, nucleosome positioning, sequence word composition.

INTRODUCTION

Nucleosomes are the basic unit of DNA packaging in eukaryotes and consist of a segment of DNA wound around a histone protein core Fig. (**1**). Often being compared to thread wrapped around a spool, nucleosomes are the fundamental building blocks of chromosomes. Actually, 75-90% [1-3] of eukaryotic genomic DNA is wrapped around by nucleosomes.

Nucleosome DNA, 165 bp long in *Saccharomyces cerevisiae* [4], can be divided into the core and the linker DNA. The core DNA, an invariable length of 147 bp of double-stranded DNA, is sharply bent and tightly wrapped around a disc-shaped histone protein octamer in 1.65 turns of a left-handed superhelix [5, 6]. The histone octamer is composed of two copies each having four core histone proteins H2A, H2B, H3, and H4 [5, 6]. The linker histone H1 is associated with the linker DNA as well as with the nucleosome core particle itself [5, 6]. The length of linker DNA varies with species and cell types during the cell differentiation and gene activation [5-7]. It is about 18 bp in *Saccharomyces cerevisiae* [5] and 38 bp in human [8].

*Address correspondence to this author at the State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, 197 Rui Jin Road II, Shanghai, 200025, China;
Tel: +86 21 63852639; Fax: +86 21 64678976;
E-mail: xykong@sibs.ac.cn

Packaging DNA into nucleosomes affects sequence accessibility as opposed to the linear naked DNA *in vivo* [2, 9-13]. This implies that nucleosome has fundamental influence on important DNA-dependent processes [14-17], such as DNA replication [18], gene transcription [19-21], DNA damage and repair [13], and DNA recombination in eukaryotic cells. Nucleosome is critical for gene regulation [2, 14, 22-28]. It can not only repress gene expression [29-33] but also facilitate gene transcription [34-36]. Therefore, a complete understanding of genomic control mechanisms in eukaryotes requires a detailed description of the determinants of nucleosome positioning.

Nucleosome positioning refers to the position where the DNA helix adopts with respect to the histone core [3]. The majority of nucleosomes are well-positioned with regularity along DNA sequences [12, 13, 37, 38]. Nucleosome position may be determined by DNA sequences [2, 39-41], chromatin remodelers [42, 43], ionic strength [44], and several other factors [44-47]. However, the relative importance of these factors is difficult to estimate *in vivo* [40, 48, 49], and the rules underlying these positioning effects are not well understood [7, 50]. Several studies have provided evidence of sequence-dependent manner for nucleosome positioning [39, 51-54]. Some results indicate that the intrinsic DNA sequence has a dominant role in determining the position of nucleosomes *in vivo* [41, 50, 55-58]. It has been suggested that as much as 50% of *in vivo* nucleosome positions in *Sac-*

0929-8665/12 $58.00+.00

charomyces is governed solely by the intrinsic genome DNA sequence [2]. Among multiple factors that are involved in determining the nucleosome positions, the underlying DNA sequence structure is essential [4].

Several DNA sequence motifs have been studied to search for the signals for nucleosome-positioning at the primary DNA sequence level [4, 40, 58-62]. Nevertheless, we are still not able to fully understand what the exact DNA sequence determinants are. It is anticipated that the power for genome-wide screening of the role of sequence-based nucleosome positioning is enhanced [63]. Previous limitation comes from the lack of large-scale experimental data with high-resolution. The purpose of this study is to use the genome sequence of *Saccharomyces cerevisiae* [18] to identify the nucleosome positions. This can increase the probability to detect nucleosome positioning signals. A number of studies have been performed in an attempt to determine nucleosome positioning signals at the level of DNA sequence using different computational methods, such as matched mirror position filter (MMPF) [64], and structure-based new developed scoring functions [65]. The present study was initiated in an attempt to use the method of minimum redundancy maximum relevance (mRMR) feature selection to identify the most important sequence features that either favor or inhibit nucleosome positioning.

MATERIALS AND METHODS

Data Preparation

The sequence reads for the H3/H4-containing nucleosomes were mapped by Mavrich *et al.* [38]. *Saccharomyces cerevisiae* genome sequences and genomic nucleosome distributions data were downloaded from Pugh's team website. We analyzed the data by combining W- and C-strand datasets. Positions for 53,021 consensus nucleosome core particles were identified by at least three sequencing reads each with the length being greater than 100 bp (see the Supplementary Material S1 and S2 for details). The sites between nucleosomes core particles were defined as linker locations, and 50,299 linker DNA sequences with at least 6bp were identified (see the Supplementary Material S3 and S4 for details). The 147 bp nucleosome-forming-related core DNA sequences were assigned as positive samples, while the nucleosome-inhibiting-related linker DNA sequences between 6bp and 2581bp were assigned as negative samples. We represent each sequence using the frequency of each overlapping *k*-mers, where *k*=1 to 6 (A, T, G, C, AA, AT, AG, AC, TA, TT, etc.). It is different to Reynolds's work [59] which limits the set of *k*-mers to length 1, 2, and 3. In this paper, the *k*-mers were called sequence words, or the words for abbreviation. Thus, each sequence is converted into a fixed-length (5,460 for exactly) vector of word frequencies and is labeled 1, 2 for core and linker sequences, respectively. Finally, we constructed a matrix (with sequences as row entries and with words as column entries), with the frequencies as its elements, as the input for mRMR feature selection as described below.

MRMR Method

MRMR was originally developed by Peng *et al*. [66]. It ranks each feature based on both its relevance to the target and the redundancy between features. A "good" feature is characterized by the maximum relevance to the target variable or by the minimum redundancy within the feature. Both relevance and redundancy are defined by mutual information (MI), which estimates how much one vector is related to another. MI is defined as follows:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (1)$$

where x and y are two vectors; $p(x,y)$ is the joint probability density; $p(x)$ and $p(y)$ are the marginal probability densities for *x* and *y*, respectively. The detailed description for the mRMR algorithm has been elaborated in [67-69], and hence isn't further repeated.

NEAREST NEIGHBOR ALGORITHM

Nearest neighbor algorithm (NNA) is used to classify the location of a DNA sequence at the nucleosome or linker. NNA makes its decision by calculating the similarities between the test sample and all the training samples. Different distance scales can be applied for this purpose, such as the Euclidean distance [70], the Hamming distance [71], and the Mahalanobis distance [72]. Here, the similarity between vectors $\mathbf{P}_x$ and $\mathbf{P}_y$ is defined by [73]

$$D(\mathbf{P}_x,\mathbf{P}_y) = 1 - \frac{\mathbf{P}_x \cdot \mathbf{P}_y}{\|\mathbf{P}_x\| \cdot \|\mathbf{P}_y\|} \quad (2)$$

where $\mathbf{P}_x \cdot \mathbf{P}_y$ is the inner product of $\mathbf{P}_x$ and $\mathbf{P}_y$, and $\|\mathbf{P}\|$ represents the module of vector $\mathbf{P}$. The smaller $D(\mathbf{P}_x,\mathbf{P}_y)$ is, the more similar $\mathbf{P}_x$ to $\mathbf{P}_y$ is. In NNA, given a query vector $\mathbf{P}_x$ and a training set P = {P_1, P_2, …, P_i, …, P_N}, P_x will be designated to the same class as its nearest neighbor P_i in **P**. In other words, if

$$\mu = \arg\min_i D(\mathbf{P}_x,\mathbf{P}_i) \quad (3)$$

where μ is the argument of *i* that minimizes $D(\mathbf{P}_x,\mathbf{P}_i)$, and if $\mathbf{P}_i$ belongs to the μ-th class, then the query vector $\mathbf{P}_x$ is assigned to the same class. For a detailed description for NNA, refer to [74].

Feature Selection

The mRMR step is used to determine which features are better and more important than others. The next step is to determine how many and which features should be selected. Here we use the incremental feature selection (IFS) method to solve this problem.

In the mRMR step, we obtain *N* feature sets from the ordered feature set *S*, with the *i*-th set being

$$S_i = \{f_0, f_1, ..., f_i\} \quad (0 \le i \le N-1) \quad (4)$$

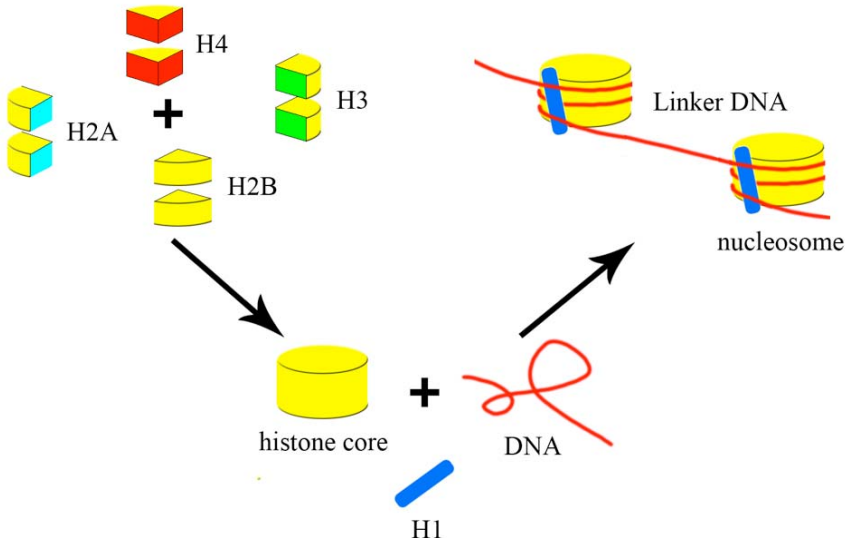

Fig. (1). A schematic illustration to show that nucleosomes are the basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound around a histone protein core.

For each i $(0 < i < N-1)$, an NNA predictor would be constructed with the feature set S_i. And jackknife test [74] was adopted to check its prediction accuracy. Finally, we can obtain an IFS curve with index i as the x-axis and the overall accuracy as the y-axis. The feature set $S_{\text{optimal}} = \{f_0, f_1, \ldots, f_h\}$ would be seen as the optimal feature set if the point in IFS curve with h as the x-axis is the highest in overall accuracy. The reason for selecting the jackknife test to examine the prediction quality is as follows. In the literature, the following three cross-validation methods are usually used to assess a statistical predictor for its anticipated accuracy: independent dataset test, sub-sampling (K-fold cross-validation) test, and jackknife test [71]. However, as elucidated by [75] and demonstrated by Eq.1 in [76], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset. This method has been increasingly used and widely recognized by investigators to examine the quality of various predictors [67, 68, 77-107].

The following indices in statistics are often used in the literature to evaluate the prediction quality:

$$\begin{cases} \Re_{\text{sen}} = \dfrac{\text{TP}}{\text{TP+FN}} \\ \Re_{\text{spe}} = \dfrac{\text{TN}}{\text{TN+FP}} \\ \Re_{\text{acc}} = \dfrac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \\ \Re_{\text{pos}} = \dfrac{\text{TP}}{\text{TP+FP}} \end{cases} \tag{5}$$

where $\Re_{\text{sen}}$, $\Re_{\text{spe}}$, $\Re_{\text{acc}}$ and $\Re_{\text{pos}}$ reflect the sensitivity, specificity, accuracy, and the positive prediction rate, respectively; TP and TN are the true positive and negative probabilities, respectively; and FP and FN are the false positive and negative probabilities, respectively (see Fig. (**2**)).

STATISTICAL ANALYSIS METHODS

By combining mRMR and IFS methods, we can obtain the optimal feature set, with the most important words affecting nucleosome formation. Furthermore, we used the point biserial correlation coefficient [108] to classify each of these words as either the nucleosome-forming-related one or the nucleosome-inhibiting-related one. Instead of calculating the correlation between two continuous variables, we calculated the correlation between a binary variable *via* the point biserial correlation coefficients using the following equation:

$$\gamma_{pq} = \frac{Y_p - Y_q}{S_y}\sqrt{pq} \tag{6}$$

where S_y is the standard deviation of all the continuous variable, and p, q are the proportions of the two values of the binary variables respectively. All the continuous variables are split into two parts based on their corresponding binary variables, and Y_p, Y_q represent the average value of the continuous variables in the two parts, respectively.

In this study, the point biserial correlation coefficients between the frequencies of each word and the type of samples (positive or negative ones) were calculated. The frequency $f_i = n_i/N$, where n_i is the number of the word's copies, and N is the total number of all words available in this sample. T-test is also used to see whether there are significant differences between the feature's frequencies in the two types of samples. If the feature's point biserial correlation coefficient is significantly greater than 0 (p-value < 0.05 in t-test), it means the frequency of this word is positively related to nucleosomes formation. On the other hand, if the point biserial correlation is significantly smaller than 0 (p-value

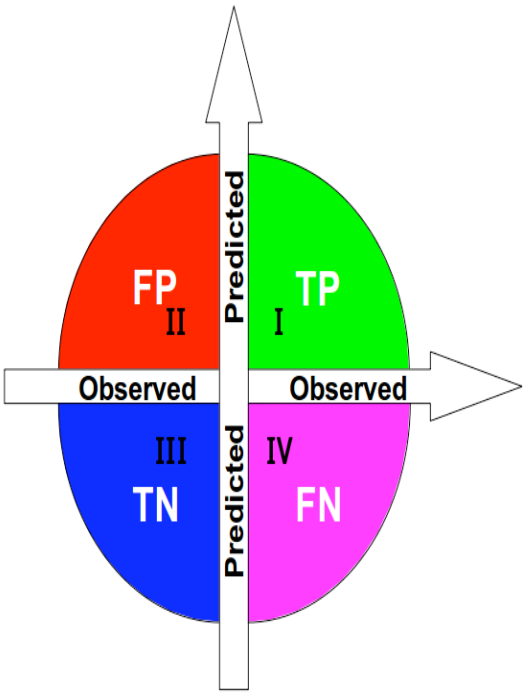

Fig. (2). An illustration to show (I) TP (true positive) quadrant (green) for correct prediction of positive dataset, (II) FP (false positive) quadrant (red) for incorrect prediction of negative dataset; (III) TN (true negative) quadrant (blue) for correct prediction of negative dataset; and (IV) FN (false negative) quadrant (pink) for incorrect prediction of positive dataset.

< 0.05), this word is regarded as a feature negatively related to the appearance of nucleosomes, i.e., nucleosome inhibiting.

In addition, to find out the factors that can change the effects of a word, we tested the independence between the contents of A+T or G+C in words and the types of words, using Pearson's Chi-squared test with Yates' continuity correction. All statistical analyses were done in R language [109], including the calculations of point biserial correlation coefficients, the *t*-test, and the Pearson's Chi-squared test with Yates' continuity correction.

RESULTS

MRMR Results

The mRMR analysis is the first step in the procedure of feature selection. We chose the parameter $t = 1$ to discretize our data to three categorical states according to the equation $mean \pm (t \cdot std)$ (*mean* is the mean value and *std* is the standard deviation). The output of mRMR (for details, see Supplementary Material S5) is a table called mRMR list recording the feature indices in Eq.4. Meanwhile, mRMR also generated another table called MaxRel list to indicate the relevance of all features with the class variable. However, here only the mRMR list was needed for the feature selection procedure.

IFS Results

Each DNA sequence was represented by a vector with 5,460 dimensions, each of which represents the corresponding word frequency. In the IFS procedure, 5,460 feature sets based on the ordered feature set *S* obtained from the mRMR list were built, and 5,460 candidate models were constructed. For efficiency, we did not test all these models. Instead, we first tested the models with feature set S_0, S_{10}, S_{20}, …, S_{480} and S_{490}. The detailed results of the 1st IFS were given in Table **1**. Fig. (**3**) showed the IFS curve, displaying that of these 50 models, the one with feature set S_{30} had the best performance: 0.7650 for the overall accuracy, 0.7400 and 0.7913 for the sensitivity and specificity, respectively. Subsequently, the candidate models that need to be considered should be around S_{30}, i.e., S_{20}, S_{21}, S_{22}, …, S_{39}. The detailed results of the 2nd IFS were given in Table **2**. Fig. (**4**) showed the obtained IFS curve. The highest overall accuracy in the IFS analysis was 0.7653 with 32 features. The corresponding sensitivity and specificity were 0.7376 and 0.7944, respectively. The feature set S_{32} was considered to be the optimal feature set. Table **3** listed the optimal features obtained in the IFS procedure.

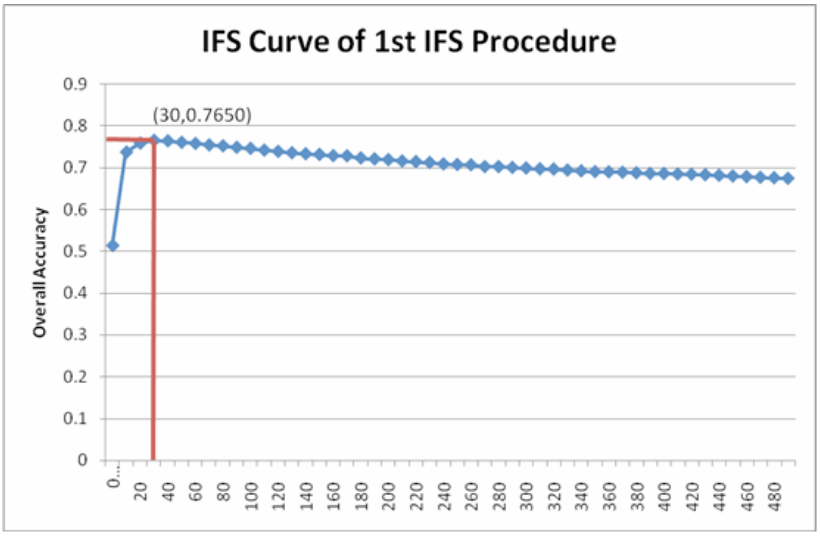

Fig. (3). The IFS curve of the first IFS procedure.

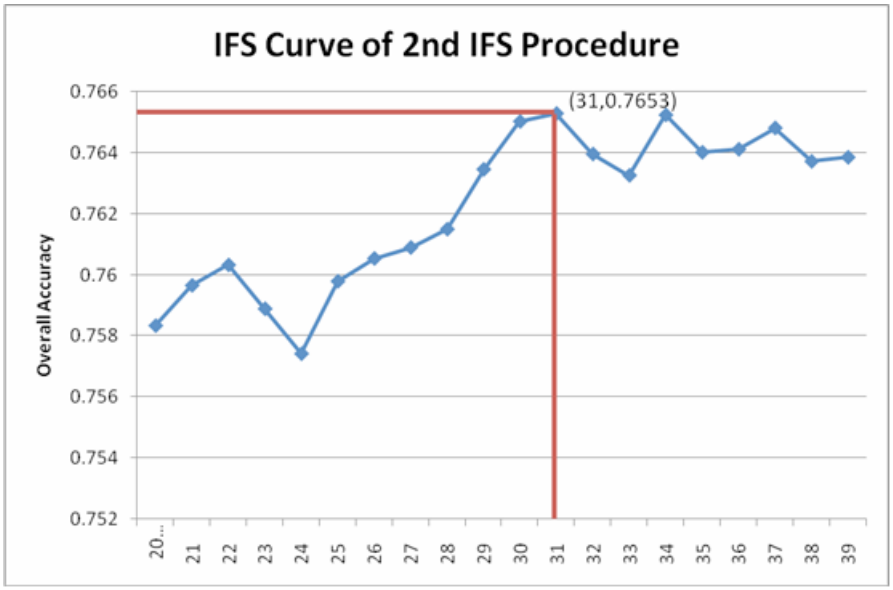

Fig. (4). The IFS curve of the second IFS procedure.

Table 1. The Results of the 1st IFS Procedure. The 50 Models with the Feature Sets S_0, S_{10}, S_{20}, …, S_{480} and S_{490} Were Tested

Model_Index	Overall_Accuracy	Sensitivity	Specificity
0	0.513173	1	0
10	0.736885	0.7567	0.715998
20	0.758333	0.760774	0.755761
30	0.765012	0.740009	0.791368
40	0.76387	0.717206	0.813058
50	0.760269	0.697554	0.826378
60	0.757569	0.681428	0.83783
70	0.754046	0.667283	0.845504
80	0.751955	0.654175	0.855027
90	0.748297	0.638294	0.864252
100	0.745858	0.628298	0.869779
110	0.741657	0.614474	0.875723
120	0.738647	0.603591	0.881012
130	0.735366	0.593557	0.884849
140	0.732956	0.58524	0.888666
150	0.731349	0.576092	0.895008
160	0.728368	0.569058	0.8963
170	0.727884	0.565663	0.898885
180	0.7229	0.552611	0.902404
190	0.72048	0.545784	0.90463

(Table 1) Contd….

Model_Index	Overall_Accuracy	Sensitivity	Specificity
200	0.718844	0.539541	0.907851
210	0.715679	0.531808	0.909501
220	0.714102	0.527357	0.910953
230	0.711914	0.52036	0.913835
240	0.708517	0.512797	0.914829
250	0.707143	0.509308	0.915684
260	0.706456	0.506158	0.917593
270	0.702139	0.497539	0.917812
280	0.702091	0.495728	0.919621
290	0.700184	0.490881	0.920814
300	0.698703	0.487995	0.920814
310	0.696893	0.483808	0.921509
320	0.696341	0.481262	0.92306
330	0.693815	0.475661	0.923776
340	0.691812	0.470851	0.92473
350	0.690147	0.466777	0.925605
360	0.689654	0.465815	0.925605
370	0.688734	0.461798	0.927951
380	0.687369	0.458196	0.928945
390	0.685985	0.455291	0.929164
400	0.685724	0.454443	0.929521
410	0.684543	0.452142	0.929521
420	0.683972	0.449445	0.931192
430	0.682685	0.445937	0.932245
440	0.681349	0.443051	0.932543
450	0.679752	0.439788	0.932702
460	0.678146	0.436544	0.932822
470	0.676432	0.432338	0.933736
480	0.675513	0.430452	0.933836
490	0.674119	0.42768	0.933895

Table 2. The Results of the 2nd IFS Procedure. The 20 Models with the Feature Sets S_{20}, S_{21}, S_{22}, …, S_{39} Were Tested

Model_Index	Overall_Accuracy	Sensitivity	Specificity
20	0.758333	0.760774	0.755761
21	0.75965	0.757945	0.761447
22	0.760317	0.755833	0.765045
23	0.758885	0.750269	0.767968
24	0.757414	0.746308	0.769121
25	0.759785	0.747176	0.773077
26	0.76053	0.746025	0.775821
27	0.760889	0.74412	0.778564
28	0.761489	0.741687	0.782362
29	0.763444	0.741876	0.786179
30	0.765012	0.740009	0.791368
31	0.765273	0.737632	0.794409
32	0.763947	0.733257	0.796298
33	0.76325	0.731107	0.797133
34	0.765225	0.731333	0.80095
35	0.764005	0.727165	0.802839
36	0.764102	0.724826	0.805503
37	0.764799	0.724732	0.807034
38	0.763705	0.721299	0.808406
39	0.76384	0.719602	0.810473

Table 3. The Features Responsible for Distinguishing the Nucleosome Forming from the Nucleosome Inhibiting Sequences

Order	Word	Score	Nucleosome Forming(+) or Inhibiting(-)
1	CG	0.181	+
2	TTAA	0.022	-
3	CAA	0.062	+
4	GG	0.032	+
5	CCTG	0.024	+
6	TTG	0.042	+
7	GC	0.028	+
8	AAAAT	0.022	-
9	CC	0.035	+
10	GAA	0.029	-
11	ATT	0.022	-

(Table 3) Contd.…

Order	Word	Score	Nucleosome Forming(+) or Inhibiting(-)
12	GCTC	0.021	+
13	TTC	0.022	-
14	CAGG	0.022	+
15	AAT	0.024	-
16	GAGC	0.021	+
17	GTGC	0.021	+
18	TTA	0.021	-
19	GCAC	0.020	+
20	GGCT	0.021	+
21	AGCC	0.021	+
22	TAA	0.022	-
23	CTAG	0.021	+
24	AC	0.023	+
25	GT	0.022	+
26	GACC	0.022	+
27	GTCC	0.021	+
28	TA	0.021	-
29	GCCT	0.021	+
30	CGGT	0.022	+
31	GGAC	0.022	+
32	TTT	0.022	-

DISCUSSION

We classified 32 words into nucleosome forming and nucleosome inhibiting features (Table **3**) by analyzing the point biserial correlation coefficients (r_{pb}) and *t*-tests. The corresponding results were shown in Table **4** and the numbers of A+T and G+C for the two classes were counted (Table **5**). The independence test, using Pearson's Chi-squared test with Yates' continuity correction, showed that correlations between the two classes were highly significant (*p*-value=2.34×10^{-9}). Both the contents of A+T and G+C are related to nucleosome forming and inhibiting. The AT-rich sequences highly inhibit nucleosome formation, while the GC-rich sequences favor nucleosome formation. This is consistent with previous studies [4, 40]. The differences between the sequences preferring to nucleosomes formation and the sequences inhibiting nucleosomes formation may lie in the DNA flexibility. Sequence-dependent DNA flexibility has been suggested to play an important role in positioning nucleosome core particles [1]. The flexible sequences would be more easily to wrap around the core histones than the rigid ones. It is well established that the inherently flexible DNA sequences can direct nucleosome assembly [56]. According to Packer's study [110]: CG, GC and GG/CC are flexible; AT and AA/TT are rigid; and TA has context-dependent flexibility. This provides a sound explanation of why a high A+T content tends to inhibit while a high G+C content tends to favor the formation of nucleosome.

Like the findings of Peckham [4], all of the 10 features related to nucleosome exclusion are the transformations of A+T content. This further confirms that DNA rigidity has a role in nucleosome of exclusion [12]. Of the 22 features related to nucleosome formation, most are the transformations of G+C content. This is due to the lower energetic cost associated with translational movements of GC-rich sequences [26]. But there are exceptions for the cases of CAA/TTG and AC/GT. This might be due to the following reason: although having little role in nucleosome positioning, they might play an important role in discriminating nucleosome forming and inhibiting sequences with other features together. Our method has made it possible to collectively consider all these features so as to lead to the best distinction of the two groups of sequences, as shown by the compelling results in Table **4**.

A study on tetranucleotide structure [111] has shown that the dinucleotides AA/TT, AT, and TA are context independ-

Table 4. The Words Related to Nucleosome Forming or Inhibiting by Ranking Point Biserial Correlation Coefficients(r_{pb})

Nucleosome Forming(+)				Nucleosome Inhibiting(-)			
Order	**Word**	r_{pb}	***p*-value**	**Order**	**Word**	r_{pb}	***p*-Value**
7	GC	0.1335	0	28	TA	-0.1952	0
9	CC	0.0936	0	11	ATT	-0.1448	0
12	GCTC	0.0925	0	15	AAT	-0.1447	0
4	GG	0.0899	0	18	TTA	-0.1434	0
5	CCTG	0.0875	0	22	TAA	-0.1422	0
16	GAGC	0.0858	0	32	TTT	-0.1116	0
14	CAGG	0.0845	0	2	TTAA	-0.0639	0
21	AGCC	0.0810	0	8	AAAAT	-0.0506	0
17	GTGC	0.0781	0	13	TTC	-0.0252	0
19	GCAC	0.0775	0	10	GAA	-0.0224	0
20	GGCT	0.0767	0				
29	GCCT	0.0765	0				
1	CG	0.0744	0				
30	CGGT	0.0734	0				
26	GACC	0.0719	0				
27	GTCC	0.0665	0				
31	GGAC	0.0621	0				
23	CTAG	0.0604	0				
24	AC	0.0572	0				
25	GT	0.0563	0				
3	CAA	0.0031	0.3189				
6	TTG	0.0022	0.4702				

Table 5. Contingency Table for Independence Test

	A/T	**G/C**	**Total**
Nucleosome forming	21	53	74
Nucleosome inhibiting	30	2	32
Total	51	55	106

ent, while CC/GG, CG, and GC are strongly context dependent. There are 22 features related to nucleosome formation (Table **3**) in which tetranucleotides consist of the majority (14/22=63.6%), while dinucleotides and trinucleotides consist of the majority (8/10=80%) in 10 features related to nucleosome exclusion. Thus, dinucleotides that inhibit nucleosome formation are generally rigid regardless of their context, while those that favor nucleosome formation are flexible with their structure depending on their tetranucleotide context. This is also fully consistent with Peckham's work [4].

In the top 32 features, there are more words related to nucleosome formation (22 words, 74 nucleotides) than to nucleosome exclusion (10 words, 32 nucleotides). This is inconsistent with Yuan's work [50] which showed that most of the top ranked sequence features appear to be related to nucleosome exclusion rather than formation. The difference

may be because their methods only consider dinucleotides but we analyze more features. The binding sites of most transcription factors (TF) are short and the degenerate sequences which occur frequently in the genome by chance [112]. Accordingly, many more matches to the known transcription factor binding sites (TFBS) may occur in the genome than previously thought [112]. Our results support the notion [25] that the genome sequence facilitates the rapid nucleosome reassembly instead of nucleosome depletion. This may be used to address the question partly why there are fewer functional TFBS than potential TFBS if nucleosomes control the binding activity of TFs by providing differential access to DNA binding sites. Strong evidence [11, 12] exists for nucleosomes regulating the accessibility of potential transcription factor binding sites. Thus, nucleosome positioning is a global determinant for the transcription factor access [12].

Up to 81% of the *Saccharomyces cerevisiae* genome DNA are organized into nucleosomes [40] and approximately 70% of nucleosomes in yeast are well positioned [12, 113, 114]. Linker regions between nucleosomes are often short (<50 bp) [40]. In our data, however, the average length of 53,005 linker DNA sequences was 80.59 (=4271586/ 53005) bp, and the base pair percent of nucleosome sequences was 64.6% (=7794087/ (7794087+4271586)), indicating that several percents of the genome, which are actually nucleosome sequences, were wrongly treated as linker sequences. Undoubtedly, this would affect the predicted results. It is anticipated that it will certainly improve our predicted results with high-resolution data available in future.

Here we used *Saccharomyces cerevisiae* in our study for its simple and the high quality and large scale of the experimental data. But the mRMR can be used in mammalian systems too, *Homo sapiens*, *Mus musculus*, *Rattus rattus*, and so on. It only needs a high-resolution nucleosome positioning data on a large scale and the genome sequence.

CONCLUSION

A feature selection method called mRMR combined with the incremental feature selection (IFS) method was applied to a benchmark dataset of 53,021 nucleosome DNA sequences and 50,299 linker DNA sequences. Different from other approaches, the mRMR method can find the motifs with minimum redundancy and maximum relevance. As a result, 32 important features were abstracted from the 5,460 features. The performance of our method achieves the overall success rate of 76.5%. Moreover, the inherent mechanism of these features to nucleosome positioning was analyzed. The findings thus obtained may provide useful insights and hints for in-depth analyzing nucleosome positioning signals and predicting the positions of nucleosome.

ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program of China (No. 2011CB510102).

SUPPORTIVE/SUPPLEMENTARY MATERIAL

S1. Genomic nucleosome sites. It shows the chromosome that each nucleosome is located in as well as the start and end position of each nucleosome.

S2. Genomic nucleosome sequences. It shows all of the *Saccharomyces cerevisiae* genomic DNA sequences in nucleosomes.

S3. Genomic linker sites. It shows the positions of all linkers between nucleosomes. It is similar to S1, showing the chromosome as well as the start and end position of each linker.

S4. Genomic linker sequences. It shows the genomic DNA sequences of all linkers between nucleosomes.

S5. mRMR analysis output. It shows the MaxRel list and mRMR list.

REFERENCES

[1] Widlund, H.R.; Kuduvalli, P.N.; Bengtsson, M.; Cao, H.; Tullius, T.D.; Kubista, M. Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J. Biol. Chem.*, **1999**, *274*, 31847-31852.

[2] Segal, E.; Fondufe-Mittendorf, Y.; Chen, L.; Thastrom, A.; Field, Y.; Moore, I.K.; Wang, J.P.; Widom, J. A genomic code for nucleosome positioning. *Nature*, **2006**, *442*, 772-778.

[3] Liu, H.; Wu, J.; Xie, J.; Yang, X.; Lu, Z.; Sun, X. Characteristics of nucleosome core DNA and their applications in predicting nucleosome positions. *Biophys. J.*, **2008**, *94*, 4597-4604.

[4] Peckham, H.E.; Thurman, R.E.; Fu, Y.; Stamatoyannopoulos, J.A.; Noble, W.S.; Struhl, K.; Weng, Z. Nucleosome positioning signals in genomic DNA. *Genome Res.*, **2007**, *17*, 1170-1177.

[5] Watson, J.D.; Baker, T.A.; Bell, S.P.; Gann, A.; Levine, M.; Losick, R. *Molecular Biology of the Gene*, 5th ed.; Benjamin Cummings, **2003**.

[6] Lewin, B. *Gene VIII*; Pearson Prentice Hall, **2004**.

[7] Luger, K. In *Encyclopedia of Life Sciences*, **2000**.

[8] Fu, Y.; Sinha, M.; Peterson, C.L.; Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.*, **2008**, *4*, e1000138.

[9] Almer, A.; Rudolph, H.; Hinnen, A.; Horz, W. Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *EMBO J.*, **1986**, *5*, 2689-2696.

[10] Mai, X.; Chou, S.; Struhl, K. Preferential accessibility of the yeast his3 promoter is determined by a general property of the DNA sequence, not by specific elements. *Mol. Cell Biol.*, **2000**, *20*, 6668-6676.

[11] Sekinger, E.A.; Moqtaderi, Z.; Struhl, K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell*, **2005**, *18*, 735-748.

[12] Yuan, G.C.; Liu, Y.J.; Dion, M.F.; Slack, M.D.; Wu, L.F.; Altschuler, S.J.; Rando, O.J. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, **2005**, *309*, 626-630.

[13] Chen, K.; Meng, Q.; Ma, L.; Liu, Q.; Tang, P.; Chiu, C.; Hu, S.; Yu, J. A novel DNA sequence periodicity decodes nucleosome positioning. *Nucleic Acids Res.*, **2008**, *36*, 6228-6236.

[14] Kornberg, R.D.; Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **1999**, *98*, 285-294.

[15] Felsenfeld, G.; Groudine, M. Controlling the double helix. *Nature*, **2003**, *421*, 448-453.

[16] Ehrenhofer-Murray, A.E. Chromatin dynamics at DNA replication, transcription and repair. *Eur. J. Biochem.*, **2004**, *271*, 2335-2349.

[17] Khorasanizadeh, S. The nucleosome: from genomic organization to genomic regulation. *Cell*, **2004**, *116*, 259-272.

[18] Mavrich, T.N.; Jiang, C.; Ioshikhes, I.P.; Li, X.; Venters, B.J.; Zanton, S.J.; Tomsho, L.P.; Qi, J.; Glaser, R.L.; Schuster, S.C.; Gilmour, D.S.; Albert, I.; Pugh, B.F. Nucleosome organization in the Drosophila genome. *Nature*, **2008**, *453*, 358-362.

[19] Grunstein, M. Nucleosomes: regulators of transcription. *Trends Genet*, **1990**, *6*, 395-400.

[20] Bondarenko, V.A.; Steele, L.M.; Ujvari, A.; Gaykalova, D.A.; Kulaeva, O.I.; Polikanov, Y.S.; Luse, D.S.; Studitsky, V.M.

Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. *Mol. Cell*, **2006**, *24*, 469-479.

[21] Whitehouse, I.; Rando, O.J.; Delrow, J.; Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, **2007**, *450*, 1031-1035.

[22] Wolffe, A.P. Transcription: in tune with the histones. *Cell*, **1994**, *77*, 13-16.

[23] Workman, J.L.; Kingston, R.E. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.*, **1998**, *67*, 545-579.

[24] Wyrick, J.J.; Holstege, F.C.; Jennings, E.G.; Causton, H.C.; Shore, D.; Grunstein, M.; Lander, E.S.; Young, R.A. Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature*, **1999**, *402*, 418-421.

[25] Lee, C.K.; Shibata, Y.; Rao, B.; Strahl, B.D.; Lieb, J.D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.*, **2004**, *36*, 900-905.

[26] Chung, H.R.; Vingron, M. Sequence-dependent Nucleosome Positioning. *J. Mol. Biol.*, **2008**.

[27] Lam, F.H.; Steger, D.J.; O'Shea, E.K. Chromatin decouples promoter threshold from dynamic range. *Nature*, **2008**, *453*, 246-250.

[28] Tirosh, I.; Barkai, N. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.*, **2008**, *18*, 1084-1091.

[29] Knezetic, J.A.; Luse, D.S. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II *in vitro*. *Cell*, **1986**, *45*, 95-104.

[30] Lorch, Y.; LaPointe, J.W.; Kornberg, R.D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, **1987**, *49*, 203-210.

[31] Han, M.; Grunstein, M. Nucleosome loss activates yeast downstream promoters *in vivo*. *Cell*, **1988**, *55*, 1137-1145.

[32] Straka, C.; Horz, W. A functional role for nucleosomes in the repression of a yeast promoter. *EMBO J.*, **1991**, *10*, 361-368.

[33] Boeger, H.; Griesenbeck, J.; Kornberg, R.D. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell*, **2008**, *133*, 716-726.

[34] McPherson, C.E.; Shim, E.Y.; Friedman, D.S.; Zaret, K.S. An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. *Cell*, **1993**, *75*, 387-398.

[35] Schild, C.; Claret, F.X.; Wahli, W.; Wolffe, A.P. A nucleosome-dependent static loop potentiates estrogen-regulated transcription from the Xenopus vitellogenin B1 promoter *in vitro*. *EMBO J.*, **1993**, *12*, 423-433.

[36] Stunkel, W.; Kober, I.; Seifart, K.H. A nucleosome positioned in the distal promoter region activates transcription of the human U6 gene. *Mol. Cell Biol.*, **1997**, *17*, 4397-4405.

[37] Loden, M.; van Steensel, B. Whole-genome views of chromatin structure. *Chromosome Res.*, **2005**, *13*, 289-298.

[38] Mavrich, T.N.; Ioshikhes, I.P.; Venters, B.J.; Jiang, C.; Tomsho, L.P.; Qi, J.; Schuster, S.C.; Albert, I.; Pugh, B.F. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **2008**, *18*, 1073-1083.

[39] Ioshikhes, I.P.; Albert, I.; Zanton, S.J.; Pugh, B.F. Nucleosome positions predicted through comparative genomics. *Nat. Genet.*, **2006**, *38*, 1210-1215.

[40] Lee, W.; Tillo, D.; Bray, N.; Morse, R.H.; Davis, R.W.; Hughes, T.R.; Nislow, C. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **2007**, *39*, 1235-1244.

[41] Kaplan, N.; Moore, I.K.; Fondufe-Mittendorf, Y.; Gossett, A.J.; Tillo, D.; Field, Y.; Leproust, E.M.; Hughes, T.R.; Lieb, J.D.; Widom, J.; Segal, E. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **2008**, *17*, 17.

[42] Vignali, M.; Hassan, A.H.; Neely, K.E.; Workman, J.L. ATP-dependent chromatin-remodeling complexes. *Mol. Cell Biol.*, **2000**, *20*, 1899-1910.

[43] Narlikar, G.J.; Fan, H.Y.; Kingston, R.E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, **2002**, *108*, 475-487.

[44] Pusarla, R.H.; Vinayachandran, V.; Bhargava, P. Nucleosome positioning in relation to nucleosome spacing and DNA sequence-specific binding of a protein. *FEBS J.*, **2007**, *274*, 2396-2410.

[45] Orphanides, G.; LeRoy, G.; Chang, C.H.; Luse, D.S.; Reinberg, D. FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell*, **1998**, *92*, 105-116.

[46] Fitzgerald, D.J.; Anderson, J.N. DNA distortion as a factor in nucleosome positioning. *J. Mol. Biol.*, **1999**, *293*, 477-491.

[47] Schwabish, M.A.; Struhl, K. Asf1 mediates histone eviction and deposition during elongation by RNA polymerase II. *Mol. Cell*, **2006**, *22*, 415-422.

[48] Albert, I.; Mavrich, T.N.; Tomsho, L. P.; Qi, J.; Zanton, S. J.; Schuster, S. C.; Pugh, B. F. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, **2007**, *446*, 572-576.

[49] Shivaswamy, S.; Bhinge, A.; Zhao, Y.; Jones, S.; Hirst, M.; Iyer, V.R. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.*, **2008**, *6*, e65.

[50] Yuan, G.C.; Liu, J.S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, **2008**, *4*, e13.

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

[52] Simpson, R.T. Nucleosome positioning *in vivo* and *in vitro*. *Bioessays*, **1986**, *4*, 172-176.

[53] Lowary, P.T.; Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **1998**, *276*, 19-42.

[54] Boyle, A.P.; Davis, S.; Shulha, H.P.; Meltzer, P.; Margulies, E.H.; Weng, Z.; Furey, T.S.; Crawford, G.E. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **2008**, *132*, 311-322.

[55] Liu, K.; Stein, A. DNA sequence encodes information for nucleosome array formation. *J. Mol. Biol.*, **1997**, *270*, 559-573.

[56] Bailey, K.A.; Reeve, J.N. DNA repeats and archaeal nucleosome positioning. *Res. Microbiol.*, **1999**, *150*, 701-709.

[57] Shen, C.H.; Clark, D.J. DNA sequence plays a major role in determining nucleosome positions in yeast CUP1 chromatin. *J. Biol. Chem.*, **2001**, *276*, 35209-35216.

[58] Fernandez, A.G.; Anderson, J.N. Nucleosome positioning determinants. *J. Mol. Biol.*, **2007**, *371*, 649-668.

[59] Reynolds, S.M.; Bilmes, J.A.; Noble, W.S. Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens. *PLoS Comput. Biol.*, **2010**, *6*, e1000834.

[60] Collings, C.K.; Fernandez, A.G.; Pitschka, C.G.; Hawkins, T.B.; Anderson, J.N. Oligonucleotide sequence motifs as nucleosome positioning signals. *PLoS ONE*, **2010**, *5*, e10933.

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

[62] Cohanim, A.B.; Kashi, Y.; Trifonov, E.N. Yeast nucleosome DNA pattern: deconvolution from genome sequences of S. cerevisiae. *J. Biomol. Struct. Dyn.*, **2005**, *22*, 687-694.

[63] Segal, M.R. Re-cracking the nucleosome positioning code. *Stat. Appl. Genet. Mol. Biol.*, **2008**, *7*, Article14.

[64] Wu, Q.; Wang, J.; Yan, H. Prediction of nucleosome positions in the yeast genome based on matched mirror position filtering. *Bioinformation*, **2009**, *3*, 454-459.

[65] Cui, F.; Zhurkin, V.B. Structure-based analysis of DNA sequence patterns guiding nucleosome positioning *in vitro*. *J. Biomol. Struct. Dyn.*, **2010**, *27*, 821-841.

[66] Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2005**, *27*, 1226-1238.

[67] He, Z.S.; Zhang, J.; Shi, X.H.; Hu, L.L.; Kong, X.G.; Cai, Y.D.; Chou, K.C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, **2010**, *5*, e9603.

[68] Huang, T.; Shi, X.H.; Wang, P.; He, Z.; Feng, K.Y.; Hu, L.; Kong, X.; Li, Y.X.; Cai, Y.D.; Chou, K.C. Analysis and Prediction of the Metabolic Stability of Proteins Based on Their Sequential Features, Subcellular Locations and Interaction Networks. *PLoS ONE*, **2010**, *5*, e10972.

[69] Chen, L.; Feng, K.Y.; Cai, Y.D.; Chou, K.C.; Li, H.P. Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition. *BMC Bioinformatics*, **2010**, *11*, 293.

[70] Chou, K.C.; Zhang, C.T. A correlation coefficient method to predicting protein structural classes from amino acid compositions. *Eur. J. Biochem.*, **1992**, *207*, 429-433.

[71] Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol*., **1995**, *30*, 275-349.

[72] Chou, K.C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins: Struct. Funct. Genet.*, **1995**, *21*, 319-344.

[73] Chou, K.C.; Cai, Y.D. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun*., **2003**, *311*, 743-747.

[74] Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[75] Chou, K.C.; Shen, H.B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **2008**, *3*, 153-162.

[76] Chou, K.C.; Shen, H.B. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci.*, **2010**, *2*, 1090-1103.

[77] Chen, C.; Chen, L.X.; Zou, X.Y.; Cai, P.X. Predicting protein structural class based on multi-features fusion. *J. Theor. Biol*., **2008**, *253*, 388-392.

[78] Du, P.; Li, Y. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. *J. Theor. Biol*., **2008**, *253*, 579-589.

[79] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol*., **2008**, *252*, 350-356.

[80] Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *J. Theor. Biol*., **2009**, *261*, 449-458.

[81] Wang, T.; Xia, T.; Hu, X.M. Geometry preserving projections algorithm for predicting membrane protein types. *J. Theor. Biol*., **2010**, *262*, 208-213.

[82] Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z.; Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol*., **2009**, *259*, 366–372.

[83] Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.*, **2007**, *248*, 546–551.

[84] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.*, **2009**, *16*, 27-31.

[85] Ding, H.; Luo, L.; Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept. Lett*., **2009**, *16*, 351-355.

[86] Jiang, X.; Wei, R.; Zhang, T.L.; Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein Pept. Lett.*, **2008**, *15*, 392-396.

[87] Joshi, R.R.; Sekharan, S. Characteristic peptides of protein secondary structural motifs. *Protein Pept. Lett.*, **2010**, *17*, 1198-1206.

[88] Li, F.M.; Li, Q.Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Pept. Lett.*, **2008**, *15*, 612-616.

[89] Li, S.; Li, H.; Li, M.; Shyr, Y.; Xie, L.; Li, Y. Improved prediction of lysine acetylation by support vector machines. *Protein Pept. Lett.*, **2009**, *16*, 977-983.

[90] Lin, H.; Ding, H.; Guo, F.B.; Zhang, A.Y.; Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **2008**, *15*, 739-744.

[91] Lin, Z.H.; Wang, H.L.; Zhu, B.; Wang, Y.Q.; Lin, Y.; Wu, Y.Z. Estimation of Affinity of HLA-A*0201 Restricted CTL Epitope Based on the SCORE Function. *Protein Pept. Lett*., **2009**, *16*, 561-569.

[92] Liu, T.; Zheng, X.; Wang, C.; Wang, J. Prediction of Subcellular Location of Apoptosis Proteins using Pseudo Amino Acid Composition: An Approach from Auto Covariance Transformation. *Protein Pept. Lett.*, **2010**, *17*, 1263-1269.

[93] Lu, J.; Niu, B.; Liu, L.; Lu, W.C.; Cai, Y.D. Prediction of small molecules' metabolic pathways based on functional group composition. *Protein Pept. Lett*., **2009**, *16*, 969-976.

[94] Mohabatkar, H. Prediction of Cyclin Proteins Using Chou's Pseudo Amino Acid Composition. *Protein Pept. Lett.*, **2010**, *17*, 1207-1214.

[95] Nanni, L.; Lumini, A. A Further Step Toward an Optimal Ensemble of Classifiers for Peptide Classification, a Case Study: HIV Protease. *Protein Pept. Lett.*, **2009**, *16*, 163-167.

[96] Shi, M.G.; Huang, D.S.; Li, X.L. A Protein Interaction Network Analysis for Yeast Integral Membrane Protein. *Protein Pept. Lett.*, **2008**, *15*, 692-699.

[97] Tian, F.; Lv, F.; Zhou, P.; Yang, Q.; Jalbout, A.F. Toward prediction of binding affinities between the MHC protein and its peptide ligands using quantitative structure-activity relationship approach. *Protein Pept. Lett.*, **2008**, *15*, 1033-1043.

[98] Yang, X.Y.; Shi, X.H.; Meng, X.; Li, X.L.; Lin, K.; Qian, Z.L.; Feng, K.Y.; Kong, X.Y.; Cai, Y.D. Classification of transcription factors using protein primary structure. *Protein Pept. Lett.*, **2010**, *17*, 899-908.

[99] Gao, Q.B.; Jin, Z.C.; Ye, X.F.; Wu, C.; He, J. Prediction of nuclear receptors with optimal pseudo amino acid composition. *Anal. Biochem.*, **2009**, *387*, 54-59.

[100] Qiu, J.D.; Huang, J.H.; Liang, R.P.; Lu, X.Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. *Anal. Biochem*., **2009**, *390*, 68-73.

[101] Xiao, X.; Wang, P.; Chou, K.C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput.Chem.*, **2009**, *30*, 1414-1423.

[102] Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput.Chem.*, **2010**, 10.1002/jcc.21616.

[103] Xiao, X.; Lin, W.Z.; Chou, K.C. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *J. Comput.Chem.*, **2008**, *29*, 2018-2024.

[104] Xiao, X.; Shao, S.H.; Huang, Z.D.; Chou, K.C. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J. Comput.Chem.*, **2006**, *27*, 478-482.

[105] Chou, K.C.; Shen, H.B. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE*, **2010**, *5*, e9931.

[106] Chou, K.C.; Shen, H.B. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE*, **2010**, *5*, e11335.

[107] Wang, Y.C.; Wang, X.B.; Yang, Z.X.; Deng, N.Y. Prediction of Enzyme Subfamily Class *via* Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature. *Protein Pept. Lett.*, **2010**, *17*, 1441-1449.

[108] Edwards, A.L. *An Introduction to Linear Regression and Correlation*; W.H. Freeman: San Francisco, **1976**.

[109] R-Development-Core-Team *R: A language and environment for statistical computing*; R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, **2008**.

[110] Packer, M.J.; Dauncey, M.P.; Hunter, C.A. Sequence-dependent DNA structure: dinucleotide conformational maps. *J. Mol. Biol*., **2000**, *295*, 71-83.

[111] Packer, M.J.; Dauncey, M.P.; Hunter, C.A. Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol*., **2000**, *295*, 85-103.

[112] Narlikar, L.; Gordân, R.; Hartemink, A.J. In *Research in Computational Molecular Biology*; Springer Berlin / Heidelberg, **2007**; Vol. 4453.

[113] Ercan, S.; Simpson, R.T. Global chromatin structure of 45,000 base pairs of chromosome III in a- and alpha-cell yeast and during mating-type switching. *Mol. Cell Biol*., **2004**, *24*, 10026-10035.

[114] Rando, O.J.; Ahmad, K. Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.*, **2007**, *19*, 250-256.

Received: February 21, 2011 **Revised: June 21, 2011** **Accepted: June 22, 2011**