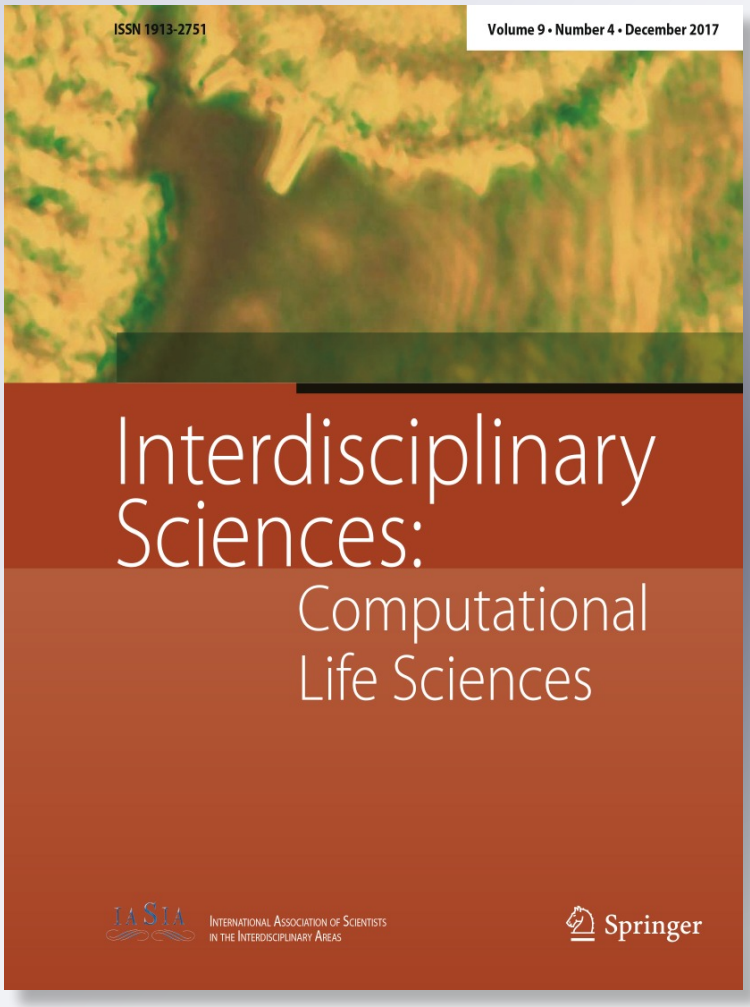
Predicting the Organelle Location of Noncoding RNAs Using Pseudo Nucleotide Compositions

Pengmian Feng, Jidong Zhang, Hua Tang, Wei Chen & Hao Lin

Interdisciplinary Sciences: Computational Life Sciences
Computational Life Sciences

ISSN 1913-2751
Volume 9
Number 4

Interdiscip Sci Comput Life Sci (2017) 9:540-544
DOI 10.1007/s12539-016-0193-4



Springer

Interdiscip Sci Comput Life Sci (2017) 9:540–544
https://doi.org/10.1007/s12539-016-0193-4

ORIGINAL RESEARCH ARTICLE

Predicting the Organelle Location of Noncoding RNAs Using Pseudo Nucleotide Compositions

Pengmian Feng[1] · Jidong Zhang[2] · Hua Tang[3] · Wei Chen[4] · Hao Lin[5]

Received: 20 May 2016 / Revised: 28 September 2016 / Accepted: 6 October 2016 / Published online: 13 October 2016

Abstract Noncoding RNAs (ncRNAs) are implicated in various biological processes. Recent findings have demonstrated that the function of ncRNAs correlates with their provenance. Therefore, the recognition of ncRNAs from different organelle genomes will be helpful to understand their molecular functions. However, the weakness of experimental techniques limits the progress toward studying organellar ncRNAs and their functional relevance. As a complement of experiments, computational method provides an important choice to identify ncRNA in different organelles. Thus, a computational model was developed to identify ncRNAs from kinetoplast and mitochondrion organelle genomes. In this model, RNA sequences are encoded by "pseudo dinucleotide composition." It was observed by the jackknife test that the overall success rate achieved by the proposed model was 90.08 %. We hope that the proposed method will be helpful in predicting ncRNA organellar locations.

Keywords Noncoding RNA · Organelle · Support vector machine · Pseudo nucleotide composition · RNA structural property

✉ Wei Chen
chenweiimu@gmail.com

✉ Hao Lin
hlin@uestc.edu.cn

[1] School of Public Health, North China University of Science and Technology, Tangshan 063000, China

[2] Department of Immunology, Zunyi Medical College, Zunyi 563000, China

[3] Department of Pathophysiology, Sichuan Medical University, Luzhou 646000, China

[4] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China

[5] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics and Center for Information in Biomedicine, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

1 Introduction

Noncoding RNAs (ncRNAs) are the major products of genome and have little or no protein-coding capabilities [1]. Although ncRNAs are the genomic "dark matter," multiple lines of evidences have demonstrated that they play critical roles in many cellular processes such as proliferation, migration, apoptosis, splicing and protein localization [2–5]. In addition, ncRNAs are also implicated in human diseases such as cardiovascular diseases and neurological disorders [6, 7].

With the discovery of ncRNAs in subcellular organelles such as kinetoplast, mitochondrion and chloroplast [8], they are now revolutionizing the concept of gene regulation in organelles. Since the function of ncRNAs correlates with their subcellular localization, correctly identifying ncRNAs from different organelles will be a great help to understand their biological roles.

Because of the low abundance of ncRNAs and the fact that ncRNAs expressed only at particular developmental stages or tissues [9], experimental methods cannot pinpoint the provenances of organellar ncRNAs. With the increase in genomic sequences, there is an urgent need to develop efficient computational tools for identifying ncRNAs from different organelle genomes, which will facilitate our understanding of the biological functions of ncRNAs.

Keeping this in mind, a support vector machine-based model was proposed to predict the organelle location of ncRNAs, in which the RNA samples were encoded using the pseudo dinucleotide composition. In the jackknife test, the proposed model obtained an accuracy of 90.08 % for identifying ncRNAs from different organelle genomes.

2 Materials and Methods

2.1 Benchmark Dataset

ncRNAs were downloaded from the database NONCODE version 4.0 [1]. According to the annotation of NONCODE, we obtained 356 ncRNAs from three kinds of organelle genomes, including 148 ncRNAs from the kinetoplast genome, 145 from the mitochondrion genome and 63 from the chloroplast genome. A high-quality benchmark dataset was built according to the following procedures: (1) To get rid of redundancy and sequence bias, the ncRNAs with high similarity were removed by using the CD-HIT software [10] with the cutoff threshold of 80 %. (2) Since the number of ncRNAs from the chloroplast genome is less than 100, for providing a significant statistics, only ncRNAs from kinetoplast and mitochondrion genomes were retained. Finally, a benchmark dataset containing 232 ncRNAs was obtained, of which 126 ncRNAs from the kinetoplast genome and 106 ncRNAs from the mitochondrion genome, which are provided in Supplementary Information S1.

2.2 RNA Sequence Formulation

Suppose a RNA sequence with L nucleic acid residues, $R_1R_2R_3 \ldots R_i \ldots R_L$, where R_i is the residue at position i and it can be adenine (A), cytosine (C), guanine (G) or uridine (U). The straightforward method to formulate the sequences is using nucleic acid composition as follows,

$$\mathbf{R} = [f(\mathrm{A}), f(\mathrm{C}), f(\mathrm{G}), f(\mathrm{U})]^{\boldsymbol{T}} \tag{1}$$

where f(A), f(C), f(G) and f(U) are the frequencies of A, C, G and U in RNA sequence, respectively. However, it missed the sequence-order information. If using the dinucleotide composition, i.e., f(AA), f(AC), f(AG), …, f(UU), although the most contiguous local sequence-order information is included, the global sequence-order information still could not be reflected.

To deal with this problem, the pseudo nucleotide composition was proposed [11] and has been widely used in computational genomics [12–19]. Recently, two flexible web servers were developed to generate pseudo nucleotide compositions [11, 20]. The pseudo nucleotide composition of RNA sequences can be defined as [11],

$$\mathbf{R} = [r_1\ r_2\ \cdots\ r_{4^k}\ r_{4^k+1}\ \cdots\ r_{4^k+\lambda}]^{\boldsymbol{T}} \tag{2}$$

where

$$r_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 4^k) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (4^k < u \le 4^k + \lambda) \end{cases} \tag{3}$$

In Eq. 3, f_u $(u = 1, 2, \cdots, 4^k)$ is the frequency of the non-overlapping k-tuple nucleotides in the RNA sequence. λ is the number of the total counted ranks of the correlations along a RNA sequence, and w is the weight factor [11]. The concrete values for λ, w and k will be given in the following: The correlation factor θ_j represents the j-tier structural correlation factor between all the jth most contiguous k-tuple nucleotide $T_i = R_iR_{i+1}\ldots R_{i+k-1}$ and is defined as,

$$\theta_j = \frac{1}{L-j-k+1}\sum_{i=1}^{L-j-k+1}\Theta(T_i, T_{i+j}) \quad (j = 1,\ 2\ ,\ \ldots,\ \lambda;\ \lambda < L) \tag{4}$$

For example, θ_1 is called the first-tier correlation factor that reflects the sequence-order correlation between all the most contiguous k-tuple nucleotide along a RNA sequence (Fig. 1a); θ_2, the second-tier correlation factor between all the second most contiguous k-tuple nucleotide (Fig. 1b); θ_3, the third-tier correlation factor between all the third most contiguous k-tuple nucleotide (Fig. 1c); and so forth. The correlation function $\Theta(T_i, T_j)$ is given by

$$\Theta(T_i, T_j) = \frac{1}{v}\sum_{u=1}^{v}\left[P_u(T_i) - P_u(T_j)\right]^2 \tag{5}$$

where v is the number of RNA physicochemical properties.

2.3 RNA Physicochemical Properties

It has been reported that RNA structures are interrelated with the functions of most ncRNAs [21, 22]. Since the structure of RNA is determined by the complex pattern of base–base interaction [23], the six RNA local structural properties (Shift, Slide, Rise, Twist, Tilt, Roll) as reported in [24] were used to define the pseudo nucleotide composition. The detailed values for the six local structural property parameters are given in Table 1.

Therefore, k is equal to 2, meaning that the pseudo dinucleotide composition (PseDNC) was used, and v is equal to 6, indicating the number of RNA physicochemical properties. $P_u(T_j)$ is the value of the uth (u = 1, 2, …, 6) property for the dinucleotide T_i at position i, and $P_u(T_j)$ is the value for the dinucleotide T_j at position j.

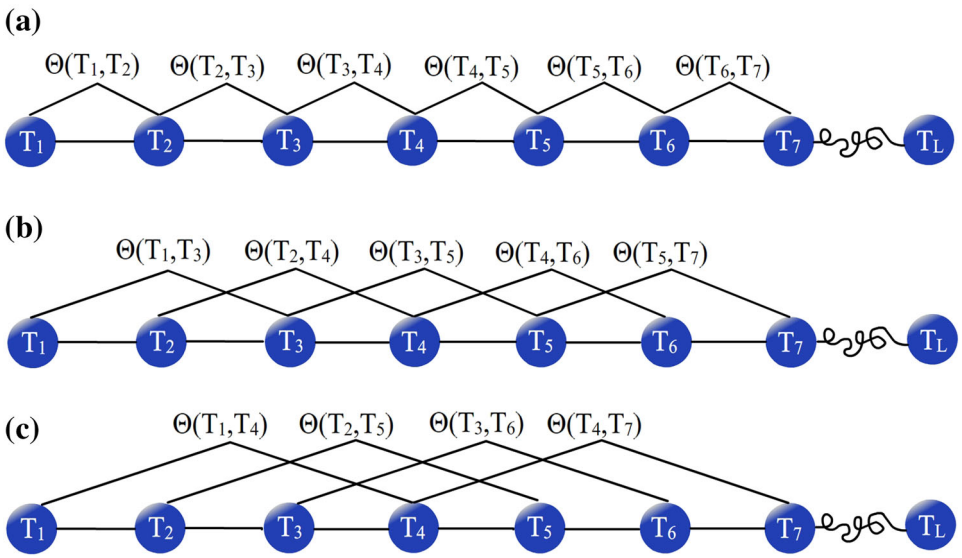


Fig. 1 Schematic illustration to show the correlations of dinucleotides along a RNA sequence. **a** The first-tier correlation reflects the sequence-order mode between all the most contiguous non-overlapping *k*-tuple nucleotide. **b** The second-tier correlation reflects the sequence-order mode between all the second most contiguous non-overlapping *k*-tuple nucleotide. **c** The third-tier correlation reflects the sequence-order mode between all the third most contiguous non-overlapping *k*-tuple nucleotide

Table 1 Concrete values of the six local structure properties for RNA dinucleotides

Dinucleotide	$P_1(T_i)$	$P_2(T_i)$	$P_3(T_i)$	$P_4(T_i)$	$P_5(T_i)$	$P_6(T_i)$
AA	−0.08	−1.27	3.18	31.0	−0.8	7.0
AC	0.23	−1.43	3.24	32.0	0.8	4.8
AG	−0.04	−1.50	3.3	30.0	0.5	8.5
AU	−0.06	−1.36	3.24	33.0	1.1	7.1
CA	0.11	−1.46	3.09	31.0	1.0	9.9
CC	−0.01	−1.78	3.32	32.0	0.3	8.7
CG	0.3	−1.89	3.3	27.0	−0.1	12.1
CU	−0.04	−1.50	3.3	30.0	0.5	8.5
GA	0.07	−1.70	3.38	32.0	1.3	9.4
GC	0.07	−1.39	3.22	35.0	0.0	6.1
GG	−0.01	−1.78	3.32	32.0	0.3	12.1
GU	0.23	−1.43	3.24	32.0	0.8	4.8
UA	−0.02	−1.45	3.26	32.0	−0.2	10.7
UC	0.07	−1.70	3.38	32.0	1.3	9.4
UG	0.11	−1.46	3.09	31.0	1.0	9.9
UU	0.08	−1.27	3.18	31.0	−0.8	7.0

In this table, the following symbols were used to represent the six physical structures of dinucleotide: P_1 for "Shift," P_2 for "Slide," P_3 for "Rise," P_4 for "Twist," P_5 for "Tilt," P_6 for "Roll." The data were obtained from [24]

Note that before substituting them into Eq. 5, all the original values $P_u(T_i)$ ($u = 1, 2, \ldots, 6$) were subjected to a standard conversion, as described by the following equation,

$$P'_u(T_i) = \frac{P_u(T_i) - <P_u(T_i)>}{\mathrm{SD}(P_u(T_i))} \tag{6}$$

where the symbol $<>$ is taking the average of the quantity therein over the 16 different dinucleotides and SD is the corresponding standard deviation.

2.4 Support Vector Machine

Support vector machine (SVM) is a classic machine learning algorithm and has been successfully used in computational genomics and proteomics [25–31]. In the current study, the LibSVM package 3.18 was used to implement SVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/∼cjlin/libsvm/, and the radial basis kernel function (RBF) was used to obtain the best classification hyperplane in the current study. In the SVM operation engine, the regularization parameter C and the kernel width parameter γ were optimized via an optimization procedure using a grid search approach defined by

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \text{with step of } 2 \\ 2^{-15} \le \gamma \le 2^{-5} & \text{with step of } 2^{-1} \end{cases}. \tag{7}$$

2.5 Performance Evaluation

Four metrics, namely sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthew's correlation coefficient (MCC), are widely used to measure the performance of a binary model [32, 33], which are expressed as

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{8}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100\% \tag{9}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \times 100\% \tag{10}$$

$$\mathrm{Mcc} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + FN) \times (TP + FP) \times (TN + FP)} \tag{11}$$

where TP represents the number of true positive, TN represents the number of true negative, FP represents the number of false positive and FN represents the number of false negative, respectively.

3 Results and Discussion

3.1 Cross-Validation

As demonstrated in Ref. [34], the jackknife test is the least arbitrary and most objective cross-validation method. Therefore, the jackknife test was used to examine the performance of the model proposed in the present work. In the jackknife test, each RNA sequence in the benchmark dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

3.2 Parameter Optimization

As shown in Eqs. 2–3, the performance of the present model depends on the two parameters w and λ. Generally speaking, the greater the λ is, the more global sequence-order information the model contains. However, if λ is too large, it would reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy due to over-fitting or "high-dimension disaster" problem [35]. Therefore, our searching for the optimal values of the two parameters is in the range of $w \in [0, 1]$ and $\lambda \in [1, 10]$ with the steps of 0.1 and 1, respectively.

In order to save the computational time, the five fold cross-validation method was used to optimize w and λ. We found that when $w = 0.1$ and $\lambda = 10$, a peak of 90.09 % was obtained for the *Acc* (Fig. 2). Accordingly, these two

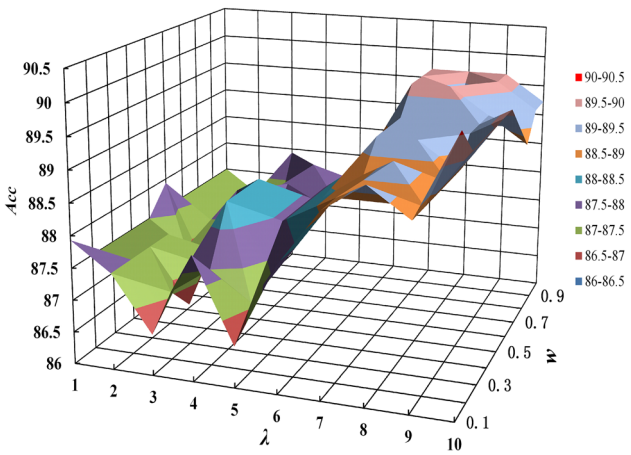


Fig. 2 3D graph to show the accuracies obtained in the fivefold cross-validation with different values of w and λ

Table 2 Comparison of different classifiers for identifying ncRNAs by the jackknife test

Classifier	Sn (%)	Sp (%)	Acc (%)	MCC
J48 Tree	85.71	88.68	87.07	0.74
RBF network	86.51	62.26	75.43	0.51
Random Forest	92.06	83.01	87.93	0.77
Naïve Bayes	23.81	94.34	56.03	0.25
SVM	91.26	88.76	90.08	0.80

numerical values, $w = 0.1$ and $\lambda = 10$, were used for the two parameters in the following analysis.

The jackknife test performance of the proposed method in identifying ncRNAs from kinetoplast and mitochondrion genomes is listed in Table 2. As given in Table 2, an accuracy of 90.08 % was obtained with the sensitivity of 91.26 %, specificity of 88.76 % and MCC of 0.80.

3.3 Comparison with Other Methods

To further testify its superiority, the predictive results of the proposed method were compared with that of other commonly used classifiers, i.e., Naïve Bayes, J48 Tree, RBF network and Random Forest as performed in WEKA [36]. The jackknife test results of different classifiers for identifying ncRNAs in the benchmark dataset are also reported in Table 2.

It is shown that the four metrics of the proposed model are all higher than that of J48 Tree and RBF network. Although the sensitivity and specificity of the proposed method are lower than those of Random Forest and Naïve Bayes, respectively, its accuracy and MCC are all higher than those of Random Forest and Naïve Bayes. These results suggest that the proposed SVM-based model can be effectively used to identify ncRNAs from different organelle genomes.

4 Conclusions

By using the pseudo dinucleotide composition to encode RNA sequences, we proposed a support vector machine-based model to discriminate ncRNAs from kinetoplast and mitochondrion genomes. The high success rates obtained from the jackknife test indicate that the model is very promising.

Although the current model is trained by the benchmark dataset only containing ncRNAs from kinetoplast and mitochondrion organelle genomes, it can also be extended to identify ncRNAs from other organelle genomes with the collections of ncRNAs from other organelles. Therefore, we hope that the proposed method could provide some novel insights into the research on organellar ncRNAs.

Compliance with Ethical Standards

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

References

1. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y (2014) NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res 42(Database issue):D98–D103. doi:10.1093/nar/gkt1222
2. Mattick JS (2011) Long noncoding RNAs in cell and developmental biology. Semin Cell Dev Biol 22(4):327. doi:10.1016/j.semcdb.2011.05.002
3. Clark MB, Mattick JS (2011) Long noncoding RNAs in cell biology. Semin Cell Dev Biol 22(4):366–376. doi:10.1016/j.semcdb.2011.01.001
4. Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136(4):629–641. doi:10.1016/j.cell.2009.02.006
5. Ma L, Bajic VB, Zhang Z (2013) On the classification of long non-coding RNAs. RNA Biol 10(6):925–933. doi:10.4161/rna.24604
6. Maass PG, Luft FC, Bahring S (2014) Long non-coding RNA in health and disease. J Mol Med 92(4):337–346. doi:10.1007/s00109-014-1131-8
7. Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. Trends Cell Biol 21(6):354–361. doi:10.1016/j.tcb.2011.04.001
8. Lung B, Zemann A, Madej MJ, Schuelke M, Techritz S, Ruf S, Bock R, Huttenhofer A (2006) Identification of small non-coding RNAs from mitochondria and chloroplasts. Nucleic Acids Res 34(14):3842–3852. doi:10.1093/nar/gkl448
9. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25(18):1915–1927. doi:10.1101/gad.17446611
10. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. doi:10.1093/bioinformatics/bts565
11. Chen W, Lei TY, Jin DC, Lin H, Chou KC (2014) PseKNC: a flexible web server for generating pseudo *k*-tuple nucleotide composition. Anal Biochem 456:53–60. doi:10.1016/j.ab.2014.04.001
12. Chen W, Feng P, Ding H, Lin H, Chou KC (2015) iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 490:26–33. doi:10.1016/j.ab.2015.08.021
13. Chen W, Lin H, Chou KC (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst 11(10):2620–2634. doi:10.1039/c5mb00155b
14. Chen W, Feng PM, Deng EZ, Lin H, Chou KC (2014) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem 462:76–83. doi:10.1016/j.ab.2014.06.022
15. Feng P, Chen W, Lin H (2014) Prediction of CpG island methylation status by integrating DNA physicochemical properties. Genomics 104(4):229–233. doi:10.1016/j.ygeno.2014.08.011
16. Chen W, Feng PM, Lin H, Chou KC (2014) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Res Int 2014:623149. doi:10.1155/2014/623149
17. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30(11):1522–1529. doi:10.1093/bioinformatics/btu083
18. Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41(6):e68. doi:10.1093/nar/gks1450
19. Feng P, Jiang N, Liu N (2014) Prediction of DNase I hypersensitive sites by using pseudo nucleotide compositions. Sci World J 2014:740506. doi:10.1155/2014/740506
20. Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou KC (2015) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31(1):119–120. doi:10.1093/bioinformatics/btu602
21. Novikova IV, Hennelly SP, Sanbonmatsu KY (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res 40(11):5034–5051. doi:10.1093/nar/gks071
22. Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, Dubois A, Sanglier-Cianferani S, Van Dorsselaer A, Clerc P, Avner P, Visvikis A, Branlant C (2010) 2-D structure of the A region of Xist RNA and its implication for PRC2 association. PLoS Biol 8(1):e1000276. doi:10.1371/journal.pbio.1000276
23. Xu XJ, Chen SJ (2015) Physics-based RNA structure prediction. Biophys Rep 1(1):2–13
24. Perez A, Noy A, Lankas F, Luque FJ, Orozco M (2004) The relative flexibility of B-DNA and A-RNA duplexes: database analysis. Nucleic Acids Res 32(20):6144–6151. doi:10.1093/nar/gkh954
25. Lin H, Liu WX, He J, Liu XH, Ding H, Chen W (2015) Predicting cancerlectins by the optimal g-gap dipeptides. Sci Rep 5:16964. doi:10.1038/srep16964
26. Ding H, Li D (2015) Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids 47(2):329–333. doi:10.1007/s00726-014-1862-4
27. Feng P, Lin H, Chen W, Zuo Y (2014) Predicting the types of J-proteins using clustered amino acids. BioMed Res Int 2014:935719. doi:10.1155/2014/935719
28. Feng PM, Chen W, Lin H, Chou KC (2013) iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem 442(1):118–125. doi:10.1016/j.ab.2013.05.024
29. Chen W, Feng P, Lin H (2012) Prediction of replication origins by calculating DNA structural properties. FEBS Lett 586(6):934–938. doi:10.1016/j.febslet.2012.02.034
30. Liu WX, Deng EZ, Chen W, Lin H (2014) Identifying the subfamilies of voltage-gated potassium channels using feature selection technique. Int J Mol Sci 15(7):12940–12951. doi:10.3390/ijms150712940
31. Lin H, Chen W, Ding H (2013) AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PLoS ONE 8(10):e75726. doi:10.1371/journal.pone.0075726
32. Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS ONE 7(10):e47843. doi:10.1371/journal.pone.0047843
33. Liu B, Fang L, Wang S, Wang X, Li H, Chou KC (2015) Identification of microRNA precursor with the degenerate *k*-tuple or Kmer strategy. J Theor Biol 385:153–159. doi:10.1016/j.jtbi.2015.08.025
34. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273(1):236–247. doi:10.1016/j.jtbi.2010.12.024
35. Wang T, Yang J, Shen HB, Chou KC (2008) Predicting membrane protein types by the LLDA algorithm. Protein Pept Lett 15(9):915–921
36. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20(15):2479–2481. doi:10.1093/bioinformatics/bth261