Data in Brief 5 (2015) 376–378

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Benchmark data for identifying $N^6$-methyladenosine sites in the *Saccharomyces cerevisiae* genome

Wei Chen [a,d,*], Pengmian Feng [b], Hui Ding [b], Hao Lin [c,d], Kuo-Chen Chou [d,e]

[a] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China*
[b] *School of Public Health, North China University of Science and Technology, Tangshan 063000, China*
[c] *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*
[d] *Gordon Life Science Institute, Belmont, MA, United States*
[e] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia*

ARTICLE INFO

*Article history:*
Received 25 August 2015
Received in revised form
30 August 2015
Accepted 10 September 2015
Available online 30 September 2015

*Keywords:*
N6-methyladenosine sites
PseKNC
PseAAC

ABSTRACT

This data article contains the benchmark dataset for training and testing iRNA-Methyl, a web-server predictor for identifying $N^6$-methyladenosine sites in RNA (Chen et al., 2015 [15]). It can also be used to develop other predictors for identifying $N^6$-methyladenosine sites in the *Saccharomyces cerevisiae* genome.



## Specifications table

| | |
|---|---|
| Subject area | Biology |
| More specific subject area | Bioinformatics, computational biology, biomedicine |
| Type of data | Text file |
| How data was acquired | Using flexible sliding window approach |
| Data format | Analyzed |
| | N/A |

* Corresponding author at: Department of Physics, Center for Genomics and Computational Biology, North China Science and Technology University, Tangshan 063000, China

http://dx.doi.org/10.1016/j.dib.2015.09.008

| | |
|---|---|
| Experimental factors | |
| Experimental features | RNA sample was formulated by combining its dinucleotide composition (DNC) [1,2] and the pseudo components [3] since nearly all the machine-learning algorithms can only handle vectors [4]. The concept of pseudo components was originally introduced to reflect the sequence patterns of protein sequences via a series of vector components [5,6] and has been widely used in computational proteomics [7]. Recently, it has been successfully extended to cover DNA [8–11] and RNA sequences [12,13] as well. For the detailed development process in this regard, see a recent review particle [14]. |
| Data source location | Chengdu 610054, China |
| Data accessibility | In Appendix A of this paper and at the web-site http://lin.uestc.edu.cn/server/iRNAMethy/data |

**Value of the data**

- N6-methyladenosine ($m^6A$) is one of the most abundant RNA methylations and plays very important roles in many biological processes [15].
- For in-depth understanding the regulatory mechanism of $m^6A$, it is indispensable to characterize its sites in a genome-wide scope.
- The data can be used to develop computational predictors or high throughput tools for identifying the $m^6A$ sites in RNA.

## 1. Background

The benchmark dataset for developing computational methods to identify the methylation sites in DNA (see, e.g., [16]) is available [17], and the information thus obtained is very useful for both basic research and drug development. But so far no existing benchmark dataset whatsoever is available for developing computational methods to identify N6-methyladenosine in RNA. The present study was initiated in an attempt to construct a benchmark dataset for the later based on the experimental observations reported by Schwartz et al. [18] recently.

## 2. Data, experimental design, materials and methods

The data presented here are the benchmark dataset for training and testing iRNA-Methyl [15] (http://lin.uestc.edu.cn/server/iRNA-Methyl), a web-server predictor for identifying $m^6A$ sites in the *S. cerevisiae* genome. By means of the $m^6A$-seq technique, Schwartz et al. [18] first identified 1,307 methylated adenine ($m^6A$) sites in the *S. cerevisiae* genome. They have observed that most of the $m^6A$ sites share a consensus motif GAC where its center base may be methylated [18]. To construct the corresponding negative benchmark dataset, we used the flexible sliding window approach [19,20] to search the *S. cerevisiae* genome, and obtained 33,280 RNA segments with exactly the same GAC consensus motif that, however, were not detected by the $m^6A$-seq technique as methylated sites. Furthermore, it had been observed via preliminary tests that when the length of the RNA segments thus derived was 51 bp, the corresponding outcomes were most promising [15]. Accordingly, the 1,307 and 33,280 RNA segments each having 51 bp long were designated as positive and negative samples, respectively. Also, since the size of the negative samples thus obtained is overwhelmingly larger than that of the positive samples, to minimize the false prediction caused by such a highly skewed benchmark dataset, we randomly picked out 1,307 RNA segments from the 33,280

negative samples to form a negative subset that has the same size with the positive one. The final benchmark dataset thus obtained contains 1,307 positive samples and 1,307 negative samples. Their detailed sequences are given in Appendix A. They can also be downloaded at the web-site http://lin.uestc.edu.cn/server/iRNAMethyl/data.

## Conflict of interest

None of the authors claims conflicting interest.

## Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (Nos. 61202256 and 61301260), and the Nature Scientific Foundation of Hebei Province (No. C2013209105).

## Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2015.09.008.

## References

[1] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.
[2] W. Chen, P.M. Feng, H. Lin, K.C. Chou, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, BioMed Res. Int. 2014 (2014) 623149.
[3] W. Chen, T.Y. Lei, D.C. Jin, K.C. Chou, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[4] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[5] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Struct. Funct. Genet. 43 (2001) 246–255 (Erratum: ibid., 2001, Vol.44, 60).
[6] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[7] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.C. Chou, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.
[8] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.
[9] P. Feng, W. Chen, H. Lin, Prediction of CpG island methylation status by integrating DNA physicochemical properties, Genomics 104 (2014) 229–233.
[10] H. Lin, E.Z. Deng, H. Ding, W. Chen, K.C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.
[11] B. Liu, F. Liu, L. Fang, X. Wang, repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects, Bioinformatics 31 (2015) 1307–1309.
[12] B. Liu, F. Liu, L. Fang, repRNA: a web server for generating various feature vectors of RNA sequences, Mol. Genet. Genom. (2015), http://dx.doi.org/10.1007/s00438-015-1078-7.
[13] B. Liu, L. Fang, F. Liu, X. Wang, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS ONE 10 (2015) e0121501.
[14] W. Chen, H. Lin, K.C. Chou, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (2015) 2620–2634.
[15] W. Chen, P. Feng, H. Ding, H. Lin, K.C. Chou, iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.
[16] Z. Liu, X. Xiao, W.R. Qiu, K.C. Chou, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77 (also Data in Brief, 2015, 4, 87–89).
[17] Z. Liu, X. Xiao, W.R. Qiu, Benchmark data for identifying DNA methylation sites via pseudo trinucleotide composition, Data Brief 4 (2015) 87–89.
[18] S. Schwartz, S.D. Agarwala, M.R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, T.S. Mikkelsen, R. Satija, G. Ruvkun, S.A. Carr, E.S. Lander, G.R. Fink, A. Regev, High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis, Cell 155 (2013) 1409–1421.
[19] K.C. Chou, Review: Prediction of protein signal sequences, Curr. Protein Pept. Sci. 3 (2002) 615–622.
[20] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. 357 (2007) 633–640.