Data in Brief 4 (2015) 87–89

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Benchmark data for identifying DNA methylation sites via pseudo trinucleotide composition

Zi Liu [a], Xuan Xiao [a,b,c,*], Wang-Ren Qiu [a], Kuo-Chen Chou [c,d]

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403 China
[b] Information School, ZheJiang Textile & Fashion College, NingBo, 315211 China
[c] Gordon Life Science Institute, Boston, MA 02478, United States
[d] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 10 March 2015
Received in revised form
29 April 2015
Accepted 29 April 2015
Available online 7 May 2015

ABSTRACT

This data article contains three benchmark datasets for training and testing iDNA-Methyl, a web-server predictor for identifying DNA methylation sites [Liu et al. *Anal. Biochem.* 474 (2015) 69–79].

Specifications table

Subject area	*Biology*
More specific subject area	*Bioinformatics and Biomedicine*
Type of data	*Text file*
How data was acquired	*Using flexible sliding window approach* [2–5]
Data format	*Analyzed*
Experimental factors	*n/a*
Experimental features	*DNA sample was formulated by combining its trinucleotide composition (TNC)* [6–8] *and the pseudo amino acid components (PseAAC)* [9–11] *of the sequence translated from the DNA sample according to its genetic codons. Meanwhile, some novel techniques in statistical analysis were introduced to train and test the predictor, such as "Neighborhood Cleaning Rule", "Synthetic Minority Over-Sampling Technique", and "Target-Jackknife Test"* [12].

DOI of original article: http://dx.doi.org/10.1016/j.ab.2014.12.009
* Corresponding author at: Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China.
E-mail addresses: xxiao@gordonlifescience.org (X. Xiao), kcchou@gordonlifescience.org (K.-C. Chou).

http://dx.doi.org/10.1016/j.dib.2015.04.021
2352-3409/

Data source location	*Jingdezhen 333403, China*
Data accessibility	With this paper and at: *http://www.jci-bioinfo.cn/DNAmethy/IDM_data.html*

1. Value of the data

- DNA methylation plays an important role in regulating a variety of biological processes and is very important for basic research and drug development as well.
- The datasets presented here are good for testing DNA methylation site identifying algorithms because of their realistic, highly unbalanced nature.
- For the first dataset (Supplementary material, File 1), users can use the original sequences to construct their own benchmark dataset, for the the 2nd dataset (Supplementary material, File 2) and the 3rd dataset (Supplementary material, File 1) users can use them to design their own predictor for identifying methylation sites.

2. Data, experimental design, materials and methods

The data presented here are three benchmark datasets for training and testing iDNA-Methyl [1] http://www.jci-bioinfo.cn/iDNA-Methyl, a web-server predictor for identifying DNA methylation sites. The DNA sample was formulated by combining its trinucleotide composition (TNC) and the pseudo amino acid components (PseAAC) of the sequence translated from the DNA sample according to its genetic codons. Sliding a window of nucleotides along each of the DNA sequences taken from MethDB (http://www.methdb.de/), and DNA sample was formulated by combining its trinucleotide composition (TNC) and the pseudo amino acid components (PseAAC) of the sequence translated from the DNA sample according to its genetic codons. In real world, the data very unbalanced. Target-jackknife was used to optimize the unbalanced benchmark dataset and minimize the consequence of this kind of mis-prediction.

I. The first dataset (Supplementary material, File 1) contains 2426 nucleotide segment samples, of which 787 are true methylation ones and 1639 are false methylation ones.
II. The 2nd dataset (Supplementary material, File 2) is the optimized benchmark dataset obtained after the NCR (Neighborhood Cleaning Rule) [13] treatments on the original benchmark dataset of the DNA methylation system. It contains 522 non-methylation samples that were removed from the negative subset, each of which corresponds to a vector with 72 components. For distinction, the real Non-methylation starts with a line of "$>$Non-Methylation code".
III. The 3rd dataset (Supplementary material, File 1) is the optimized benchmark dataset obtained after both the NCR (Neighborhood Cleaning Rule) [13] and SMOTE (Synthetic Minority Over-Sampling Technique) [14] treatments on the 1st benchmark dataset. It contains 1117 DNA methylation (including 330 hypothetical methylation created by SMOTE) and 1117 non-methylation, each of which corresponds to a vector with 72 components. For distinction, the real DNA methylation starts with a line of "$>$Methylation code" while the hypothetical DNA methylation starts with a line of "Hypothetical" [6–8].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2015.04.021.

References

[1] Z. Liu, X. Xiao, W.R. Qiu, K.C. Chou, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–79.

[2] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[3] K.C. Chou, Review: prediction of protein signal sequences, Curr. Protein Peptide Sci. 3 (2002) 615–622.
[4] K.C. Chou, H.B. Shen, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. (BBRC) 357 (2007) 633–640.
[5] H.B. Shen, K.C. Chou, Signal-3L: a 3-layer approach for predicting signal peptide, Biochem. Biophys. Res. Commun. (BBRC) 363 (2007) 297–303.
[6] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, and K.C. Chou, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Analytical Biochemistry 456 (2014) 53-60.
[7] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, and K.C. Chou, PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31 (2015) 119-120.
[8] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.C. Chou, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences Nucleic Acids Research http://dx.doi.org/10.1093/nar/gkv458 (2015).
[9] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43 (2001) 246-255.
[10] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21 (2005) 10-19.
[11] K.C. Chou, Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry 11 (2015) 218-234.
[12] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, and K.C. Chou, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. Journal of Biomolecular Structure & Dynamics (JBSD) (2014) http://dx.doi.org/10.1080/07391102.2014.998710.
[13] J. Laurikkala, Improving Identification of Difficult Small Classes by Balancing Class Distribution, Springer, Berlin Heidelberg 63–66.
[14] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2011) 321–357.