Bioinformatics, 31(20), 2015, 3362–3364
doi: 10.1093/bioinformatics/btv366
Advance Access Publication Date: 14 June 2015
Applications Note

Sequence analysis

PPUS: a web server to predict PUS-specific pseudouridine sites

Yan-Hui Li[1,*], Gaigai Zhang[2] and Qinghua Cui[3,*]

[1]Institute of Cardiovascular Sciences, Peking University Health Science Center, [2]Department of Geriatrics and Gerontology, Beijing Huaxin Hospital, the First Affiliated Hospital of Tsinghua University and [3]Department of Biomedical Informatics, Peking University Health Science Center, Beijing, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 19, 2015; revised on May 31, 2015; accepted on June 5, 2015

Abstract

Motivation: Pseudouridine (Ψ), catalyzed by pseudouridine synthase (PUS), is the most abundant RNA modification and has important cellular functions. Developing an algorithm to identify Ψ sites is an important work. And it is better if the algorithm could assign which PUS modifies the Ψ sites. Here, we developed **PPUS** (http://lyh.pkmu.cn/ppus/), the first web server to predict PUS-specific Ψ sites. **PPUS** employed support vector machine as the classifier and used nucleotides around Ψ sites as the features. Currently, **PPUS** could accurately predict new Ψ sites for PUS1, PUS4 and PUS7 in yeast and PUS4 in human. **PPUS** is well designed and friendly to user.

Availability and Implementation: Our web server is available freely for non-commercial purposes at: http://lyh.pkmu.cn/ppus/

Contact: liyanhui@bjmu.edu.cn or cuiqinghua@hsc.pku.edu.cn

1 Introduction

Pseudouridine (Ψ), the most abundant RNA modification, has important cellular functions. For example, ablation of rRNA Ψ modification by CBF5 deletion in *S. cerevisae* is lethal (Jiang *et al.*, 1993; Zebarjadian *et al.*, 1999), and mutation of human PUS1 leads to mitochondrial myopathy and sideroblastic anemia (Fujiwara and Harigae, 2013). Identifying Ψ sites is certain an important work.

Ψ was known present in different categories of non-coding RNAs such as tRNAs, rRNAs and snRNA (Ge and Yu, 2013). Recently, using newly developed Ψ-seq (Psi-seq or Pseudo-seq) methods, three works for the first time revealed that Ψ is also present in mRNA (Carlile *et al.*, 2014; Lovejoy *et al.*, 2014; Schwartz *et al.*, 2014). Ψ is catalyzed by pseudouridine synthase (PUS), and different Ψ sites are catalyzed by different PUSs (Schwartz *et al.*, 2014). However, to our knowledge, there is only one tool developed for prediction of Ψ sites in tRNA (Panwar and Raghava, 2014). No tool is available to predict Ψ sites in mRNA and other categories of RNAs, not mention predicting which PUS modifies the Ψ sites.

In this work, using PUS-specific Ψ sites found in recent works (Carlile *et al.*, 2014; Schwartz *et al.*, 2014), we developed an algorithm to predict PUS-specific Ψ sites. The algorithm could accurately predict Ψ sites modified by PUS1, PUS4 and PUS7 in yeast and PUS4 in human. Any RNA sequence can be used as input. Finally, we implemented the algorithm in our web server PPUS. PPUS is well designed and friendly to user.

2 Methods

2.1 Data source

Using newly developed Ψ-seq methods, many Ψ sites and the PUSs that modified them were identified in mRNA, tRNA, snoRNA and rRNA in recent works (Carlile *et al.*, 2014; Schwartz *et al.*, 2014). We downloaded data from both works and put all Ψ sites modified by a certain PUS together, regardless of the RNA category. To train classifiers, these Ψ sites were defined as gold standard positive samples. Corresponding to the positives, gold standard negatives were gotten by randomly choosing equal number of 'U' sites from the remaining 'U's of the same sequences. The number of gold standard positive samples for each PUS was listed in Table 1. These data could be downloaded from website: http://lyh.pkmu.cn/ppus/data.html. To avoid sampling bias, we sampled gold standard negative

Table 1. Performances of SVM on different datasets

Species	PUS	*#Pos*	*L*	*R*	*Pre*	*Rec*	*F1*	*AUC*
All-yeast		464	8	9	0.64	0.60	0.62	0.62
Yeast	PUS1	84	4	3	0.74	0.80	0.77	0.74
Yeast	PUS2	29	2	5	0.54	0.61	0.58	0.52
Yeast	PUS3	61	1	9	0.61	0.66	0.63	0.57
Yeast	PUS4	55	2	7	0.75	0.78	0.77	0.80
Yeast	PUS7	199	2	6	0.96	0.98	0.97	0.99
Yeast	CBF5	60	9	3	0.49	0.67	0.57	0.50
Human	DKC1	46	1	2	0.56	0.61	0.60	0.53
Human	PUS4	56	4	7	0.95	0.98	0.97	1.00

#pos: #positive; *Pre*: *precision*; *Rec*: *recall*; L/R: number of nucleotides at the left/right of the Ψ site; All-Yeast: all Ψ sites found in yeast; AUC: area under curve.

dataset 100 times, and then combined each negative dataset with the positives to train the classifier.

2.2 Support vector machine

The classification model for predicting Ψ sites was based on support vector machine. The software LIBSVM3.20 (Chang, 2011) was employed, in which a radial basis function (RBF) was chosen as the kernel function. The default values of parameter *c* and *g* were used. LIBSVM outputs a posterior probability for each prediction to reflect its reliability (Kwok, 1999). The larger the posterior probability is for a site, the more likely the site could be modified to Ψ. To make it easy to display in web server, the posterior probability was converted to *M*-score by Equation (1). The *M*-score is an integer between 5 and 10.

$$M\text{-score} = \text{floor}\,(\text{posterior probability} * 10) \quad (1)$$

2.3 Classifier evaluation

To evaluate performance of support vector machine (SVM), 5-fold cross validation was adopted. In each round, 20% of the samples were left out as the test set and the remaining as the training set. As in previous work (Li *et al.*, 2010), *precision*, *recall* and *F1* were used to evaluate the classifier. Of the sites predicted as Ψ sites, the numbers of true positives (*TP*) and false negatives (*FN*) were counted. Of the genes predicted as non-Ψ sites, the numbers of true negatives (*TN*) and false positives (*FP*) were also counted. Then the *precision*, *recall* and *F1* score were calculated as

$$\text{Precision} = \frac{TP}{TP+FP},\ \text{Recall} = \frac{TP}{TP+FN},\ F1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

Precision is the fraction of true positives in the predicted positives, whereas *recall* is the fraction of gold standard positives that are predicted as true positives. *F1* is used to evaluate the overall performance of a classifier. Receiver operating characteristic curve is another measure often used for classifier evaluation (Liu *et al.*, 2014a,b, 2015), thus we also compute area under curve (*AUC*) in this work. Because negative datasets were gotten by 100 random samplings, the medians of *precision*s, *recall*s, *F1*s and *AUC*s of the 100 training results were used (Table 1).

2.4 Feature selection

We used a sliding window strategy to get nucleotides around the Ψ sites as classification features. The window was in the form of '*L*Ψ*R*', where '*L*' and '*R*' represented 0–10 nt at the left and right side of Ψ site, respectively. To generate fixed window length, a dummy '*X*' nucleotide was added to sequence terminals if needed. As in previous work (Panwar and Raghava, 2014), we used binary approach and represented *A*, *C*, *G*, *U* and *X* nucleotides with {1,0,0,0,0}, {0,1,0,0,0}, {0,0,1,0,0}, {0,0,0,1,0} and {0,0,0,0,1}, respectively. To get best performance for the classifier, we optimized the window length ('*L*' and '*R*') and listed them in Table 1.

2.5 Web server

The web server was run on Linux. Programs were written in PHP (Hypertext Preprocessor). PPUS is freely available at http://lyh.pkmu.cn/ppus/.

3 Results

As described in methods, Ψ sites and the PUSs that modified them were gotten from recent works (Carlile *et al.*, 2014; Schwartz *et al.*, 2014). First, we wanted to test whether it was possible to develop an algorithm to predict Ψ sites if exclude the information of which PUS modified the sites. We put all known 464 yeast Ψ sites together as gold standard positives, and got equal number of negatives by random sampling. Unfortunately, as shown in Table 1, the median *precision, recall, F1 and AUC* were 0.64, 0.60, 0.62 and 0.62, respectively. The performance was poor, showing that if excluding PUS-specific modification information, it was hard to distinguish Ψ sites from non-Ψ sites.

Second, it was reported that PUS1, PUS4, PUS7 recognized 'AΨ', 'GUΨCNANYCY' and 'UGΨAG', respectively (Lovejoy *et al.*, 2014). Thus, we wanted to know whether it was possible to distinguish Ψ sites catalyzed by a certain PUS from non-Ψ sites. Fortunately, as shown in Table 1, the classifier performed excellent on PUS1, PUS4 and PUS7 in yeast, and PUS4 in human, with median *F1* = 0.77, 0.78, 0.98 and 0.97, respectively. However, it performed poorly on PUS2, PUS3 and CBF5 in yeast, and DKC1 in human, with median *F1* = 0.58, 0.63, 0.58 and 0.60, respectively. This might because the feature nucleotides were not found. In this work, no RNA category information was used for classifier training, so the classifiers should be RNA category insensitive. In other words, any RNA sequence can be used as input.

Third, we validated the yeast and human PUS4 classifiers with independent datasets. As shown in Table 1, there were 55 and 56 positive samples for yeast and human PUS4 synthases, respectively. To test the yeast PUS4 classifier, we used the human PUS4 Ψ sites as an independent dataset, and finally found that 52 of the 56 Ψ sites can be correctly predicted. Similarly, we used the yeast Ψ sites as an independent dataset to test the human PUS4 classifier and found that 39 of the 55 yeast Ψ sites can be correctly predicted. Both results showed that the classifiers were efficient.

In conclusion, we for the first time developed an algorithm to efficiently predict PUS-specific Ψ sites. We implemented the algorithm in our web server PPUS. We think PPUS will be an important tool for biologist in the field.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 81300253, 81422006 and 91339106) and the National High Technology Research and Development Program of China.

Conflict of Interest: none declared.

References

Carlile,T.M. *et al*. (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, **515**, 143–146.

Chang,C.-C, Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27.

Fujiwara,T. and Harigae,H. (2013) Pathophysiology and genetic mutations in congenital sideroblastic anemia. *Pediatr. Int.*, **55**, 675–679.

Ge,J. and Yu,Y.T. (2013) RNA pseudouridylation: new insights into an old modification. *Trends Biochem. Sci.*, **38**, 210–218.

Jiang,W. *et al*. (1993) An essential yeast protein, CBF5p, binds in vitro to centromeres and microtubules. *Mol. Cell. Biol.*, **13**, 4884–4893.

Kwok,J.Y. (1999) Moderating the outputs of support vector machine classifiers. *IEEE Trans. Neural. Netw.*, **10**, 1018–1031.

Li,Y.H. *et al*. (2010) Systematic analysis and prediction of longevity genes in Caenorhabditis elegans. *Mech. Ageing Dev.*, **131**, 700–709.

Liu,B. *et al*. (2014a) iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS One*, **9**, e106691.

Liu,B. *et al*. (2014b) Using distances between Top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinformatics*, **15** (Suppl. 2), S3.

Liu,B. *et al*. (2015) Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One*, **10**, e0121501.

Lovejoy,A.F. *et al*. (2014) Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PLoS One*, **9**, e110799.

Panwar,B. and Raghava,G.P. (2014) Prediction of uridine modifications in tRNA sequences. *BMC Bioinformatics*, **15**, 326.

Schwartz,S. *et al*. (2014) Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, **159**, 148–162.

Zebarjadian,Y. *et al*. (1999) Point mutations in yeast CBF5 can abolish in vivo pseudouridylation of rRNA. *Mol. Cell. Biol.*, **19**, 7461–7472.