Supplementary file : A novel RNA pseudouridine site prediction model using utility kernel and data-driven parameters*

Sourabh Patil*[1], Archana Mathur*[2][0000−0003−4522−6890], Raviprasad Aduri[1,3][0000−0001−9627−4986], and Snehanshu Saha[3,4][0000−0002−8458−604X]

[1] Dept. of Biological Sciences, BITS Pilani K K Birla Goa Campus, Goa, India
[2] Dept. of Information Science and Engineering, Nitte Meenakshi Institute of Technology, Bangalore, India
mathurarchana77@gmail.com
[3] Dept. of CS & IS and APPCAIR, BITS Pilani K K Birla Goa Campus, Goa, India
[4] HappyMonk AI, Bangalore, India
aduri@goa.bits-pilani.ac.in
scibase.snehanshu@gmail.com

1 Results: PSe-MA compared with PseUI and iRNA-PseU model; 5-fold cross validation

In K-fold CV, the data is divided into K-folds. The $k-1$ folds are used for training and the k^{th} fold is tested, and the process is repeated k times, each time with a different subset for train and test.

2 Results: PSe-MA compared with PseUI and iRNA-PseU models; 66-33% split

The 66%-33% train-test split is the standard way of CV where 66% of data is used for training and 33% of data is used for reporting the performance of the model.

*

SVM Utility kernel against PseUI and iRNA-PseU models; with 5-fold CV						
Training datasets	Model used	Sensitivity	Specificity	Accuracy	MCC	AUC
H990	iRNA-PseU (rbf)	63.21	62.34	61.32	0.43	0.65
	iRNA-PseU (linear)	65.32	64.32	63.89	0.41	0.58
	iRNA-PseU (UK)	**66.21±0.001**	**59.98±0.02**	**61.92±0.07**	**0.34±0.008**	**0.55±0.01**
	PseUI (rbf)	64.82	63.25	64.87	0.56	0.68
	PseUI (linear)	65.78	64.65	66.66	0.47	0.74
	PseUI (UK)	**70.76±0.09**	**69.54±0.07**	**71.33±0.08**	**0.46±0.05**	**0.50±0.04**
	PSe-MA (rbf)	70.30	81.62	75.96	0.52	0.76
	PSe-MA (linear)	63.03	76.97	70.00	0.41	0.70
	PSe-MA (UK)	**89.52±0.01**	**89.52±0.031**	**89.49±0.02**	**0.78±0.06**	**0.89±0.071**
S628	iRNA-PseU (rbf)	64.65	64.33	64.49	0.29	0.81
	iRNA-PseU (linear)	67.43	67.83	67.77	0.43	0.69
	iRNA-PseU (UK)	**68.55±0.06**	**69.33±0.09**	**68.61±0.04**	**0.51±0.05**	**0.69±0.02**
	PseUI (rbf)	62.10	71.02	66.56	0.33	0.69
	PseUI (linear)	64.32	71.90	65.65	0.34	0.61
	PseUI (UK)	**67.33±0.04**	**68.79±0.01**	**72.73±0.05**	**0.33±0.06**	**0.69±0.02**
	PSe-MA (rbf)	90.10	94.92	92.51	0.85	0.93
	PSe-MA (linear)	69.01	93.02	81.05	0.81	0.64
	PSe-MA (UK)	**97.69±0.07**	**97.69±0.04**	**97.61±0.01**	**0.95±0.02**	**0.97±0.05**
M994	iRNA-PseU (rbf)	72.32	61.32	65.4	0.32	0.75
	iRNA-PseU (linear)	73.31	61.83	63.07	0.48	0.65
	iRNA-PseU (UK)	**79.87±0.05**	**60.81±0.01**	**70.34±0.07**	**0.41±0.008**	**0.75±0.06**
	PseUI (rbf)	73.67	64.33	70.14	0.53	0.67
	PseUI (linear)	74.58	66.31	70.44	0.41	0.77
	PseUI (UK)	**75.43±0.08**	**67.73±0.03**	**71.32±0.06**	**0.39±0.02**	**0.79±0.01**
	PSe-MA (rbf)	78.62	77.34	77.97	0.56	0.78
	PSe-MA (linear)	75.44	74.80	77.92	0.56	0.78
	PSe-MA (UK)	**91.45±0.007**	**91.45±0.01**	**91.42±0.09**	**0.83±0.04**	**0.91±0.02**

Table 1. 5 fold Cross-validation Results-iRNA-PseU and PseUI used RBF SVM on their data; our method PSe-MA used Utility Kernel on new biological features extracted in this work; the three feature vectors are made to run on three SVM kernels viz. RBF, linear and utility kernel. Utility Kernel outperformed the other methods with a Bonferroni Comparison (BC) value of 3 on every dataset.

Different models (ML) on the three training datasets using PSe-MA						
Training datasets	Model used	Sensitivity	Specificity	Accuracy	MCC	AUC
H990	GausianNB	88.41	88.41	88.48	0.77	0.88
	Decision Trees	74.07	74.07	74.24	0.48	0.74
	Random Forest	75.97	75.97	75.96	0.52	0.75
	XGBoost	89.48	89.48	89.49	0.78	0.89
	PSe-MA (UK)	**89.52±0.001**	**89.52±0.004**	**89.49±0.02**	**0.78±0.03**	**0.89±0.07**
S628	GausianNB	94.78	94.78	94.75	0.89	0.94
	Decision Trees	85.59	85.59	85.51	0.71	0.85
	Random Forest	83.1	83.1	83.13	0.66	0.83
	XGBoost	97.04	97.04	96.98	0.93	0.97
	PSe-MA (UK)	**97.69±0.03**	**97.69±0.04**	**97.61±0.09**	**0.95±0.02**	**0.97±0.07**
M994	GausianNB	88.93	88.93	88.88	0.77	0.88
	Decision Trees	80.66	80.66	80.62	0.61	0.8
	Random Forest	77.84	77.84	77.75	0.55	0.77
	XGBoost	90.88	90.88	90.00	0.81	0.9
	PSe-MA (UK)	**91.45±0.01**	**91.45±0.03**	**91.42±0.01**	**0.83±0.10**	**0.91±0.08**

Table 2. Different ML models (5-fold CV) compared with our new set of features. Utility Kernel outperformed the other methods with a Bonferroni Comparison (BC) value of 3 on every dataset.

PSe-MA compared with PseUI and iRNA-PseU models; 66-33% split						
Training datasets	Training datasets	Sensitivity	Specificity	Accuracy	MCC	AUC
H990	iRNA-PseU (rbf)	62.41	60.23	62.44	0.32	0.56
	iRNA-PseU (linear)	60.32	61.24	63.43	0.31	0.58
	iRNA-PseU (UK)	**62.34±0.03**	**58.65±0.02**	**63.22±0.07**	**0.27±0.05**	**0.59±0.06**
	PseUI (rbf)	63.45	65.64	68.91	0.38	0.64
	PseUI (linear)	68.93	66.80	65.43	0.32	0.52
	PseUI (UK)	**71.11±0.09**	**68.34±0.03**	**70.32±0.002**	**0.44±0.04**	**0.61±0.01**
	PSe-MA (rbf)	79.81	76.21	77.32	0.42	0.71
	PSe-MA (linear)	77.78	80.61	79.20	0.79	0.58
	PSe-MA (UK)	**90.18±0.03**	**90.18±0.01**	**90.2±0.09**	**0.8±0.04**	**0.9±0.009**
S628	iRNA-PseU (rbf)	65.45	63.76	65.76	0.38	0.73
	iRNA-PseU (linear)	62.45	64.44	62.32	0.31	0.77
	iRNA-PseU (UK)	**68.55±0.07**	**69.33±0.01**	**68.61±0.06**	**0.51±0.08**	**0.69±0.02**
	PseUI (rbf)	64.32	75.02	68.62	0.54	0.54
	PseUI (linear)	64.32	72.31	67.43	0.23	0.64
	PseUI (UK)	**67.33±0.06**	**68.79±0.03**	**72.73±0.01**	**0.33±0.05**	**0.69±0.02**
	PSe-MA (rbf)	79.99	76.54	75.45	0.77	0.83
	PSe-MA (linear)	88.39	96.88	92.31	0.93	0.85
	PSe-MA (UK)	**94.18±0.08**	**94.16±0.01**	**94.17±0.07**	**0.88±0.04**	**0.94±0.001**
M994	iRNA-PseU (rbf)	72.21	65.23	68.32	0.42	0.64
	iRNA-PseU (linear)	72.45	66.54	69.82	0.41	0.76
	iRNA-PseU (UK)	**79.87±0.08**	**60.81±0.04**	**70.34±0.06**	**0.41±0.09**	**0.75±0.007**
	PseUI (rbf)	74.21	65.51	71.23	0.38	0.67
	PseUI (linear)	73.56	67.83	71.21	0.45	0.76
	PseUI (UK)	**76.58±0.01**	**69.83±0.02**	**71.32±0.04**	**0.52±0.07**	**0.65±0.05**
	PSe-MA (rbf)	78.82	79.32	75.43	0.61	0.61
	PSe-MA (linear)	77.50	80.92	79.17	0.79	0.58
	PSe-MA (UK)	**91.0±0.07**	**91.0±0.009**	**91.0±0.01**	**82.0±0.05**	**91.0±0.09**

Table 3. 66-33% Cross-validation Results:iRNA-PseU and PseUI used RBF SVM on their data; our method PSe-MA used Utility Kernel on new biological features extracted in this work; the three feature vectors are made to run on three SVM kernels viz. RBF, linear, and utility kernel. Utility Kernel outperformed the other methods with a Bonferroni Comparison (BC) value of 3 on every dataset.

Different ML models on three training datasets (66-33 % CV) using PSe-MA						
Training datasets	Training datasets	Sensitivity	Specificity	Accuracy	MCC	AUC
H990	GausianNB	88.9	88.9	88.1	0.77	0.88
	Decision Trees	77.95	77.95	77.98	0.55	0.77
	Random Forest	78.56	78.56	78.59	0.57	0.78
	XGBoost	90.1	90.1	90.18	0.8	0.89
	PSe-MA (UK)	**90.18±0.04**	**90.18±0.03**	**90.2±0.009**	**0.8±0.06**	**0.9±0.01**
S628	GausianNB	91.44	91.44	91.34	0.82	0.91
	Decision Trees	80.57	80.57	80.28	0.61	0.8
	Random Forest	84.6	84.6	84.61	0.69	0.84
	XGBoost	93.83	93.83	93.75	0.87	0.93
	PSe-MA (UK)	**94.18±0.02**	**94.16±0.04**	**94.17±0.07**	**0.88±0.09**	**0.94±0.08**
M994	GausianNB	89.1	89.1	89.1	0.78	0.89
	Decision Trees	81.71	81.71	81.41	0.63	0.81
	Random Forest	81.59	81.59	81.73	0.63	0.81
	XGBoost	89.51	89.51	89.42	0.78	0.89
	PSe-MA (UK)	**91.0±0.009**	**91.0±0.008**	**91.0±0.001**	**82.0±0.004**	**91.0±0.006**

Table 4. Different ML models (66-33% CV) compared with our new set of features. Utility Kernel outperformed the other methods with a Bonferroni Comparison (BC) value of 3 on every dataset.

3 Results on additional data

5 fold CV with different ML models on testing datasets						
Testing datasets	Model used	Sensitivity	Specificity	Accuracy	MCC	AUC
H200	Decision tress	77.82	78.92	77.79	0.74	0.84
	Random forrest	90.25	91.44	92.85	0.85	0.97
	XGBoost	91.72	90.25	93.28	0.87	0.82
	SVM rbf	74.75	77.14	75.97	0.85	0.67
	SVM linear	74.62	76.92	75.79	0.53	0.84
	SVM Utility	**95.38±0.02**	**95.38±0.005**	**95.40±0.033**	**0.90±0.009**	**0.95±0.01**
S200	Decision tress	85.24	86.75	86.00	0.86	0.72
	Random forrest	99.04	99.04	99.04	0.97	0.98
	XGBoost	98.03	98.55	98.43	0.99	0.97
	SVM rbf	96.65	96.14	96.38	0.99	0.93
	SVM linear	89.61	89.64	89.63	0.96	0.79
	SVM Utility	**99.69±0.07**	**99.69±0.002**	**99.68±0.009**	**0.99±0.02**	**0.99±0.06**

Table 5. 5 fold Cross-validation Results with a comparison of different ML models on two of the testing datasets

4 The sequence information of three Datasets- H.sapiens, S.cerevisiae and M.musculus M.

This study used three benchmark datasets for training: H990, S628, and M944. These are the same datasets used by iRNA-PseU and PseUI methods. These datasets are curated from the experimentally found ψ sites from RMBase for H. sapiens, M. musculus, and S. cerevisiae. The negative dataset is made of RNA sequences that were experimentally confirmed to not have ψ sites. In addition to the training datasets, Chen et al. [1] provided two independent testing datasets for H. sapiens and S. cerevisiae, namely H200 and S200, but not for M. musculus.

Species	Name of the training dataset	Length of the RNA sequence	Number of positive samples	Number of negative samples
H.sapiens	H-990	21	495	495
S.cerevisiae	S-628	31	314	314
M.musculus	M-994	21	472	472

Table 6. The sequence information of the three datasets

Species	Name of the training dataset	Length of the RNA sequence	Number of positive samples	Number of negative samples
H.sapiens	H-200	21	100	200
S.cerevisiae	S-200	31	100	200

Table 7. Additional Data - The sequence information of the two testing datasets

References

1. Chen, W., Tang, H., Ye, J., Lin, H., Chou, K.C.: irna-pseu: Identifying rna pseudouridine sites. Molecular Therapy. Nucleic Acids **5** (2016), https://api.semanticscholar.org/CorpusID:17740037