Electronic Supplementary Material (ESI) for Molecular Omics.

RF-Glutarysite:Random Forest based predictor for Glutarylation sites

Hussam J. AL-barakati[a], Hiroto Saigo[b], Robert H. Newman[c] & Dukka B. KC[a,*]

[a] *Department of Computational Science and Engineering, NCA&T State University, Greensboro NC 27411 , USA. E-mail: dbkc@ncat.edu*
[b] *Department of Informatics, Kyushu University, Fukuoka 819-0395, Japan.*
[c] *Department of Biology, NCA&T State University, Greensboro NC 27411.*

Supplementary Information File

Supplemental Methods:

The optimization parameters of Random Forest (RF) are:

RandomForestClassifier(bootstrap=True,class_weight=None, criterion='gini',max_depth=None,max_features='auto', max_leaf_nodes=None,min_impurity_decrease=0.0, min_impurity_split=None,min_samples_leaf=1, min_samples_split=2,min_weight_fraction_leaf=0.0, n_estimators=800,n_jobs=1,oob_score=False, random_state=0, verbose=0, warm_start=False)

Results:

Table S1. *Comparison between various machine learning algorithms based on 10-fold cross-validation using all 12,887 features. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.*

Features	ACC (%)	SN (%)	SP (%)	MCC
Random Forest (RF)	**69**	**77**	**62**	**0.39**
Support vector machine (SVM)	64	67	62	0.28
Naïve Bayes (NB)	60	66	54	0.20
K-nearest neighbor(KNN)	61	62	60	0.22

Table S2. *Comparison between various machine learning algorithms based on an independent test set using all 12,887 features. ACC: Accuracy; SN: Sensitivity; SP: Specificity; MCC: Matthew's Correlation Coefficient.*

Features	ACC (%)	SN (%)	SP (%)	MCC
Random Forest (RF)	**68**	**77**	59	**0.36**
Support vector machine (SVM)	63	55	**73**	0.27
Naïve Bayes (NB)	63	73	55	0.27
K-nearest neighbor(KNN)	63	68	59	0.27

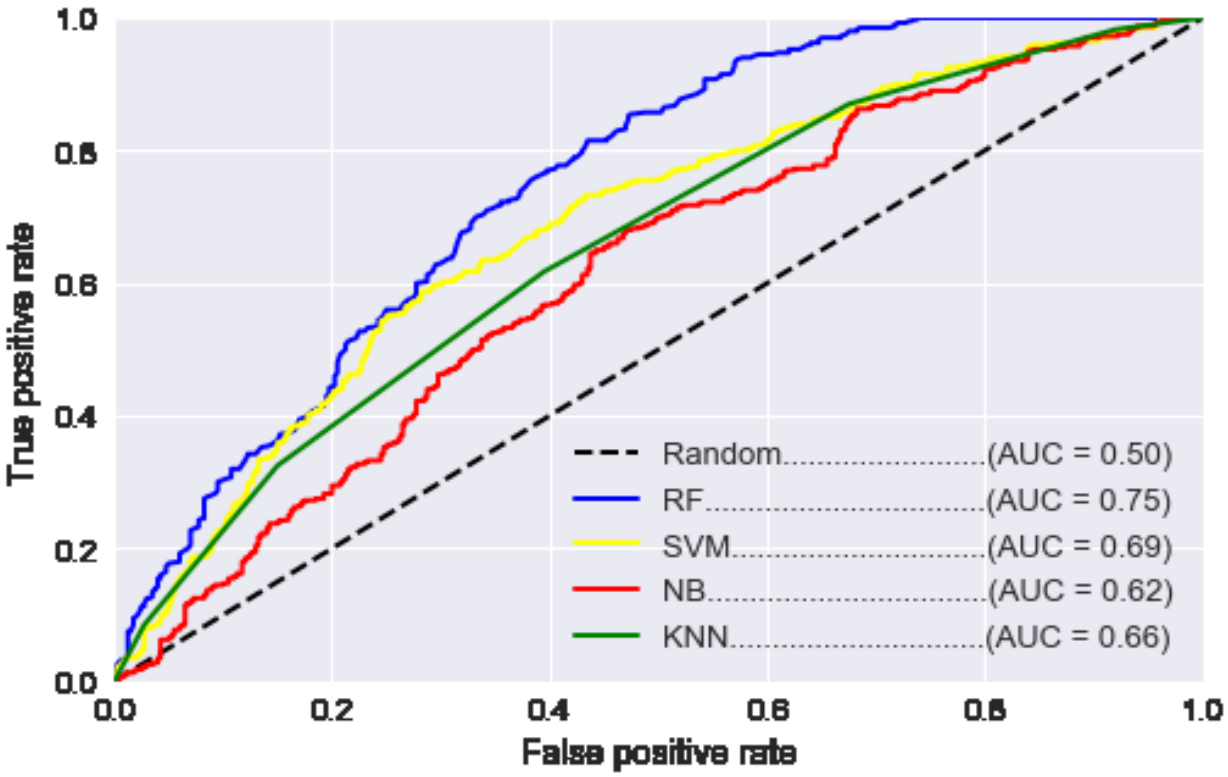

Fig. S1. *Receiver operator characteristic (ROC) curves for various machine learning algorithms using the entire feature set (12,887) based on 10-fold cross-validation. The area under the curve (AUC) for each algorithm is given in parentheses. SVM: Support vector machine; NB: Naïve Bayesian; KNN: k-nearest neighbor; RF: Random forest.*

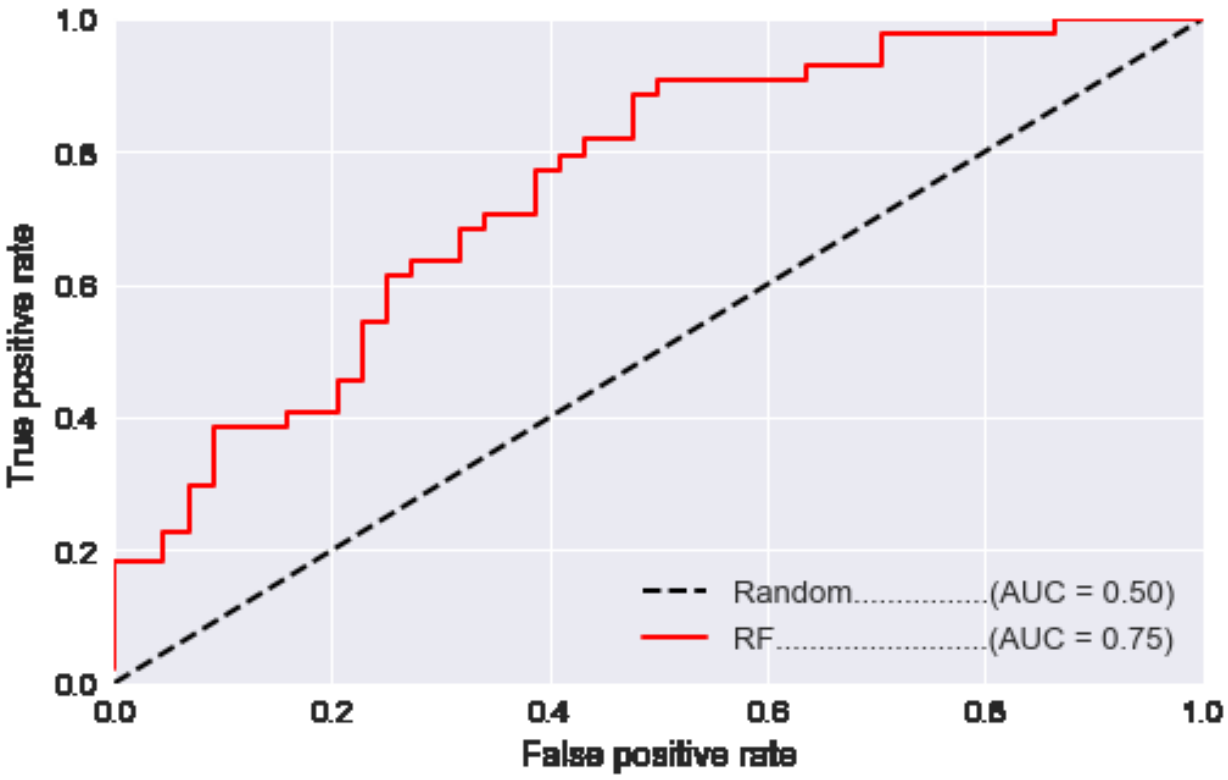

Fig. S2. *Receiver operator characteristic (ROC) curve based on the independent test set for the Random Forest (RF) classifier using the complete feature set (12,887 features).*

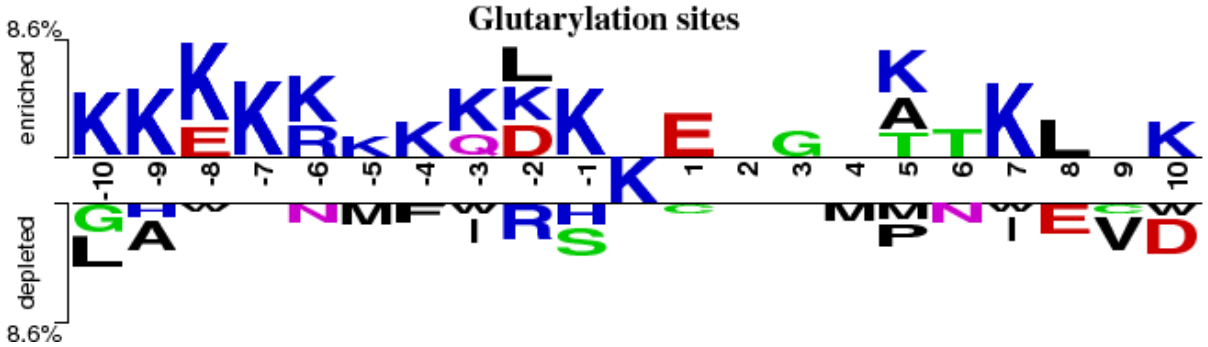

Fig. S3 *Two-sample log applied for both positive and negative sites with a threshold of P < 0.1. The logo was created using the web-based application developed by Vacic and colleagues.*[131]

To identify the optimal window size for our algorithm, we conducted a comparative analysis across multiple window sizes, feature selection thresholds and trees. These studies suggest

that window size 23 with threshold of 0.002 for feature selection and 800 trees exhibited the best performance based on MCC (for instance, please see Figs. S4 and S5, comparing MCC results for various window sizes using a threshold of 0.002 and 800 trees for 10-fold cross-validation and the independent dataset, respectively).

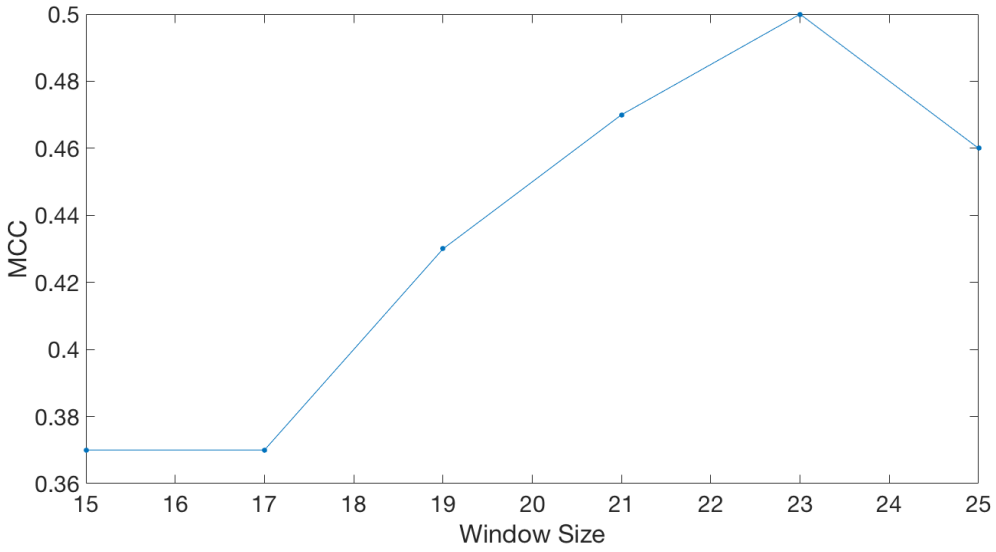

***Fig. S4**. Method performance, based on the MCC metric, for different window sizes using 10-fold cross-validation. For each window, a threshold of 0.002 and 800 trees were used for analysis based on analyses summarized in Tables S3-S8.*

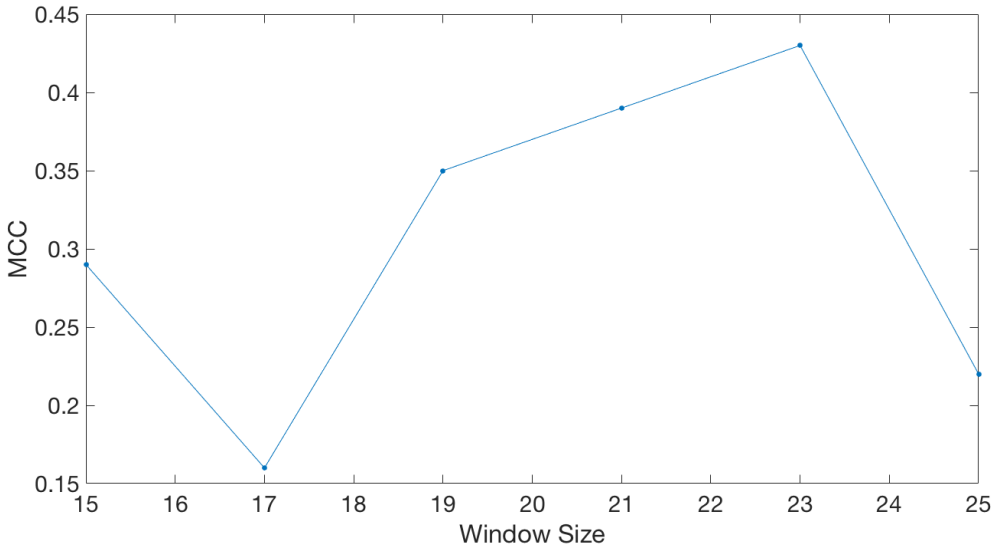

***Fig. S5**. Method performance, based on the MCC metric, for different window sizes using the independent test set. For each window size, a threshold of 0.002 and 800 trees were used for analysis based on analyses summarized in Tables S9-S14.*

The threshold and number of trees were chosen based on a series of analyses, which are summarized below:

First, we tried to find the best window size by varying the threshold of feature selection and the number of trees used for RF. We compared different windows sizes with different thresholds of feature selection with different number of trees based on both 10-fold cross-validation and the independent test. These data are summarized for each window size in Tables S3-S8 (10-fold cross-validation) and Tables S9-S14 (independent dataset).

From Tables S3 to S8, it can be seen that a window size of 23 with a threshold of 0.002 and 800 trees performed better than other windows using the same threshold (0.002) as well as other windows using different thresholds. Similar results were observed using the independent test set (Tables S9-S14).

These analyses also demonstrated that, given the same number of trees, threshold values of 0.002 or 0.003 yielded identical results. This was true using both 10-fold cross-validation and the independent test set for all windows evaluated, suggesting that further increases in the threshold does not substantially improve method performance. Therefore, the threshold was set at 0.002 for final method development.

Table S3. Comparison between windows sizes 15 using a different threshold based on 10-fold cross-validation

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (15)	307	0.001	800	66	74	58	0.33	0.73
Window-size (15)	307	0.001	400	66	73	59	0.32	0.73
Window-size (15)	307	0.001	135	65	72	59	0.31	0. 72
Window-size (15)	138	0.002	800	68	78	59	0.37	0.74
Window-size (15)	138	0.002	400	68	79	58	0.37	0.74
Window-size (15)	138	0.002	135	67	76	59	0.35	0.74
Window-size (15)	138	0.003	800	68	78	59	0.37	0.74
Window-size (15)	138	0.003	400	68	79	58	0.37	0.74
Window-size (15)	138	0.003	135	67	76	59	0.35	0.74

Table S4. Comparison between windows sizes 17 using a different threshold based on 10-fold cross-validation

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (17)	309	0.001	800	67	76	59	0.35	0.74
Window-size (17)	309	0.001	400	66	74	59	0.34	0.74
Window-size (17)	309	0.001	135	67	76	59	0.35	0.74
Window-size (17)	142	0.002	800	68	76	61	0.37	0.76
Window-size (17)	142	0.002	400	68	77	61	0.37	0.76
Window-size (17)	142	0.002	135	67	73	62	0.35	0.75
Window-size (17)	142	0.003	800	68	76	61	0.37	0.76
Window-size (17)	142	0.003	400	68	77	61	0.37	0.76
Window-size (17)	142	0.003	135	64	70	59	0.30	0.68

Table S5. Comparison between windows sizes 19 using a different threshold based on 10-fold cross-validation

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (19)	322	0.001	800	69	76	63	0.39	0.75
Window-size (19)	322	0.001	400	70	77	64	0.40	0.75
Window-size (19)	322	0.001	135	68	75	61	0.36	0.74
Window-size (19)	141	0.002	800	70	77	65	0.41	0.77
Window-size (19)	141	0.002	400	71	77	66	0.43	0.77
Window-size (19)	141	0.002	135	69	73	65	0.38	0.77
Window-size (19)	141	0.003	800	65	80	52	0.33	0.68
Window-size (19)	141	0.003	400	67	80	55	0.35	0.67
Window-size (19)	141	0.003	135	61	75	48	0.23	0.66

Table S6. Comparison between windows sizes 21 using a different threshold based on 10-fold cross-validation

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (21)	335	0.001	800	71	78	64	0.43	0.78
Window-size (21)	335	0.001	400	71	74	70	0.43	0.78
Window-size (21)	335	0.001	135	70	74	67	0.41	0.78
Window-size (21)	135	0.002	800	73	76	71	0.47	0.80
Window-size (21)	135	0.002	400	73	75	72	0.47	0.80
Window-size (21)	135	0.002	135	64	70	59	0.30	0.68
Window-size (21)	135	0.003	800	73	76	71	0.47	0.80
Window-size (21)	135	0.003	400	73	75	72	0.47	0.80
Window-size (21)	135	0.003	135	64	70	59	0.30	0.68

Table S7. Comparison between windows sizes 23 using a different threshold based on 10-fold cross-validation

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (23)	338	0.001	800	72	79	66	0.45	0.78
Window-size (23)	338	0.001	400	72	78	67	0.44	0.79
Window-size (23)	338	0.001	135	73	80	67	0.47	0.79
Window-size (23)	128	0.002	800	75	81	68	0.50	0.81
Window-size (23)	128	0.002	400	74	80	69	0.49	0.80
Window-size (23)	128	0.002	135	72	78	68	0.45	0.80
Window-size (23)	128	0.003	800	75	81	68	0.50	0.81
Window-size (23)	128	0.003	400	74	80	69	0.49	0.80
Window-size (23)	128	0.003	135	72	78	68	0.45	0.80

Table S8. Comparison between windows sizes 25 using a different threshold based on 10-fold cross-validation

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (25)	358	0.001	800	71	78	64	0.42	0.77
Window-size (25)	358	0.001	400	70	78	64	0.42	0.77
Window-size (25)	358	0.001	135	70	76	65	0.40	0.77
Window-size (25)	130	0.002	800	72	79	66	0.44	0.79
Window-size (25)	130	0.002	400	73	80	66	0.46	0.80
Window-size (25)	130	0.002	135	72	78	66	0.44	0.79
Window-size (25)	130	0.003	800	72	79	66	0.44	0.79
Window-size (25)	130	0.003	400	73	80	66	0.46	0.80
Window-size (25)	130	0.003	135	72	78	66	0.44	0.79

Table S9. Comparison between windows sizes 15 using a different threshold based on independent test set.

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (15)	307	0.001	800	62	75	50	0.25	0.61
Window-size (15)	307	0.001	400	59	70	48	0.18	0.58
Window-size (15)	307	0.001	135	54	61	48	0.16	0.58
Window-size (15)	138	0.002	800	60	80	41	0.29	0.60
Window-size (15)	138	0.002	400	63	84	43	0.39	0.60
Window-size (15)	138	0.002	135	59	70	48	0.18	0.60
Window-size (15)	138	0.003	800	60	80	41	0.29	0.60
Window-size (15)	138	0.003	400	63	84	43	0.39	0.60
Window-size (15)	138	0.003	135	59	70	48	0.18	0.60

Table S10. Comparison between windows sizes 17 using a different threshold based on independent test set.

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (17)	309	0.001	800	61	80	43	0.24	0.60
Window-size (17)	309	0.001	400	54	68	41	0.09	0.58
Window-size (17)	309	0.001	135	50	64	36	0.01	0.53
Window-size (17)	142	0.002	800	57	70	45	0.16	0.61
Window-size (17)	142	0.002	400	54	61	48	0.18	0.61
Window-size (17)	142	0.002	135	57	66	50	0.16	0.63
Window-size (17)	142	0.003	800	57	70	45	0.16	0.61
Window-size (17)	142	0.003	400	54	61	48	0.18	0.61
Window-size (17)	142	0.003	135	57	66	50	0.16	0.63

Table S11. Comparison between windows sizes 19 using a different threshold based on independent test set.

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (19)	322	0.001	800	65	80	52	0.33	0.68
Window-size (19)	322	0.001	400	65	82	50	0.33	0.69
Window-size (19)	322	0.001	135	68	82	55	0.37	0.73
Window-size (19)	141	0.002	800	67	82	52	0.35	0.68
Window-size (19)	141	0.002	400	67	80	55	0.35	0.68
Window-size (19)	141	0.002	135	61	75	48	0.23	0.68
Window-size (19)	141	0.003	800	67	82	52	0.35	0.68
Window-size (19)	141	0.003	400	67	80	55	0.35	0.68
Window-size (19)	141	0.003	135	61	75	48	0.23	0.68

Table S12. Comparison between windows sizes 21 using a different threshold based on independent test set.

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (21)	335	0.001	800	67	75	60	0.35	0.72
Window-size (21)	335	0.001	400	65	80	52	0.33	0.68
Window-size (21)	335	0.001	135	63	75	52	0.28	0.65
Window-size (21)	135	0.002	800	69	77	61	0.39	0.70
Window-size (21)	135	0.002	400	70	80	61	0.41	0.71
Window-size (21)	135	0.002	135	64	70	59	0.30	0.68
Window-size (21)	135	0.003	800	69	77	61	0.39	0.70
Window-size (21)	135	0.003	400	70	80	61	0.41	0.71
Window-size (21)	135	0.003	135	64	70	59	0.30	0.68

Table S13. Comparison between windows sizes 23 using a different threshold based on independent test set.

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (23)	338	0.001	800	69	70	68	0.38	0.77
Window-size (23)	338	0.001	400	76	77	75	0.52	0.78
Window-size (23)	338	0.001	135	72	70	75	0.45	0.78
Window-size (23)	128	0.002	800	72	73	70	0.43	0.81
Window-size (23)	128	0.002	400	70	70	70	0.41	0.80
Window-size (23)	128	0.002	135	73	77	73	0.47	0.82
Window-size (23)	128	0.003	800	72	73	70	0.43	0.81
Window-size (23)	128	0.003	400	70	70	70	0.41	0.80
Window-size (23)	128	0.003	135	73	77	73	0.47	0.82

Table S14. Comparison between windows sizes 25 using a different threshold based on independent test set.

Features	Number features	Threshold	nTree	ACC(%)	SN(%)	SP(%)	MCC	AUC
Window-size (25)	358	0.001	800	64	77	52	0.30	0.70
Window-size (25)	358	0.001	400	67	80	55	0.35	0.72
Window-size (25)	358	0.001	135	68	82	55	0.37	0.74
Window-size (25)	130	0.002	800	61	66	57	0.22	0.66
Window-size (25)	130	0.002	400	61	68	55	0.22	0.66
Window-size (25)	130	0.002	135	62	68	57	0.25	0.66
Window-size (25)	130	0.003	800	61	66	57	0.22	0.66
Window-size (25)	130	0.003	400	61	68	55	0.22	0.66
Window-size (25)	130	0.003	135	62	68	57	0.25	0.66

The comparative analysis of different window sizes, threshold values and trees outlined above suggested that a window size of 23 results in the best method performance using both 10-fold cross-validation and our independent test set. From these analyses, it is also apparent that increasing the threshold reduces the number of features (Fig. S6). Therefore, to ensure that the observed loss of features does not adversely affect method performance, we compared method performance for window size 23 using the top features for a given threshold (e.g., a threshold of 0 contained all 12,829 features while thresholds of 0.001 and 0.004 yielded 338 and 69 features, respectively). The method performance under each of these conditions is summarized in Tables S15 and S16. As can be seen in Figs. S7 and S8, 128 features consistently resulted in the best method performance using both 10-fold cross-validation (Fig. S7) and the independent test set (Fig. S8). Similar results were obtained if the number of trees remained constant (Tables S17-S18 and Figs. S9-S10). Together, these data suggest that reducing the number of features based on a defined threshold does not adversely affect model performance.

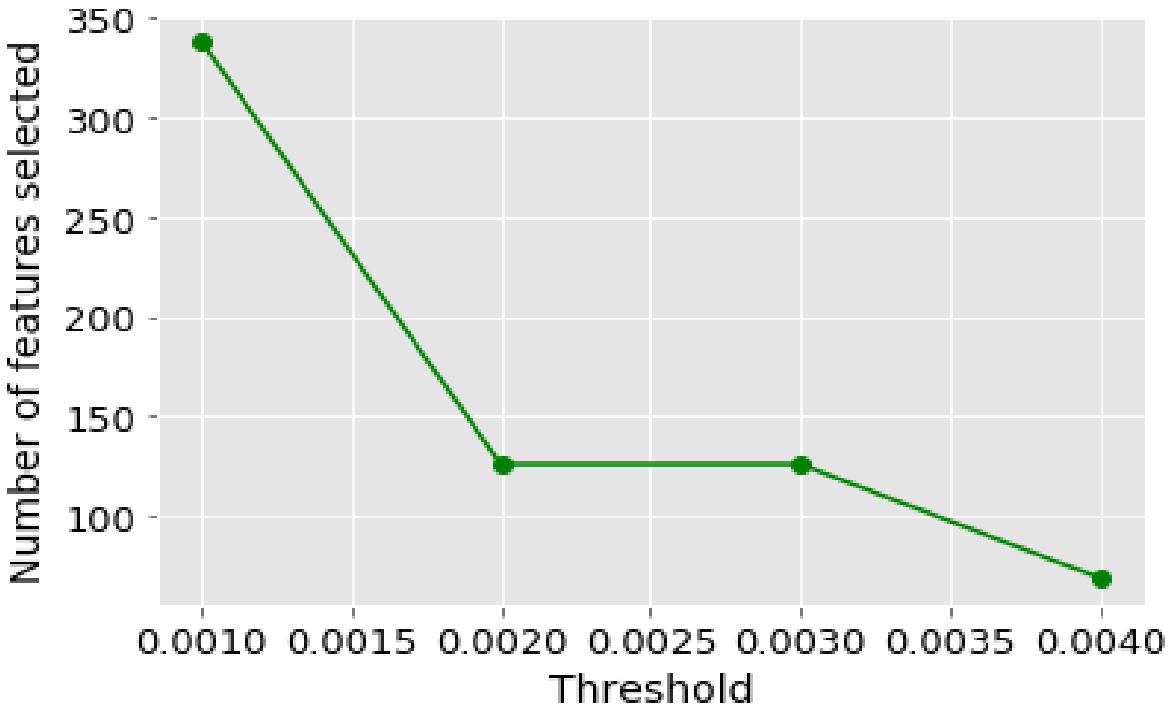

Fig. S6. *Shows different threshold with different number of feature selected*

Different trees

Table S15. Comparison between our method using different threshold based on 10-fold cross-validation.

Features	Number features	Threshold	Ntree	ACC(%)	SN(%)	SP(%)	MCC	AUC
RF-GlutarySite	12,829	0	4000	69	77	62	0.39	0.75
RF-GlutarySite	338	0.001	900	72	79	66	0.45	0.78
RF-GlutarySite	128	0.002	800	75	81	68	0.50	0.81
RF-GlutarySite	128	0.003	400	74	80	69	0.49	0.80
RF-GlutarySite	69	0.004	135	73	79	68	0.47	0.80

Table S16. Comparison between our method using different threshold based on independent test set.

Features	Number features	Threshold	Ntree	ACC(%)	SN(%)	SP(%)	MCC	AUC
RF-GlutarySite	12,887	0	4000	68	77	59	0.36	0.75
RF-GlutarySite	338	0.001	900	69	70	68	0.38	0.77
RF-GlutarySite	128	0.002	800	72	73	70	0.43	0.81
RF-GlutarySite	128	0.003	400	70	70	70	0.41	0.80
RF-GlutarySite	69	0.004	135	67	70	64	0.34	0.74

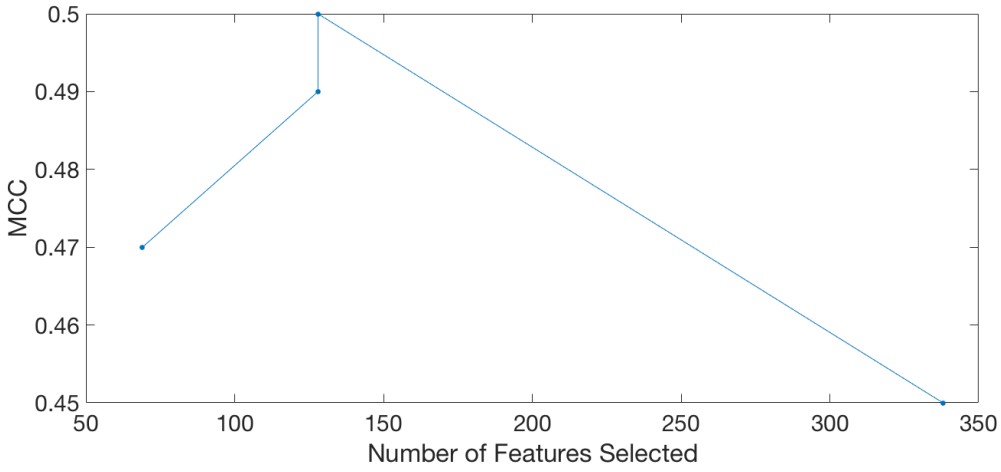

***Fig. S7**. Shows different number of feature selected against MCC metrics based on 10-fold cross validation.*

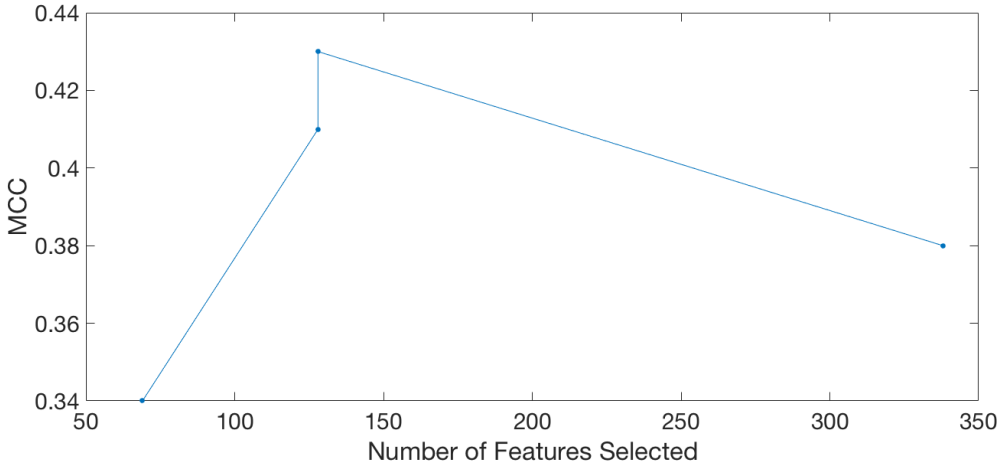

***Fig. S8**. Shows different number of feature selected against MCC metrics based on an independent set.*

Using same trees

Table S17. Comparison between our method using different threshold based on 10-fold cross-validation.

Features	Number features	Threshold	Ntree	ACC(%)	SN(%)	SP(%)	MCC	AUC
RF-GlutarySite	12,829	0	800	67	74	61	0.35	0.75
RF-GlutarySite	338	0.001	800	72	79	66	0.45	0.78
RF-GlutarySite	128	0.002	800	75	81	68	0.50	0.81
RF-GlutarySite	128	0.003	800	75	81	68	0.50	0.81
RF-GlutarySite	69	0.004	800	73	80	68	0.47	0.80

Table S18. Comparison between our method using different threshold based on independent test set.

Features	Number features	Threshold	Ntree	ACC(%)	SN(%)	SP(%)	MCC	AUC
RF-GlutarySite	12,887	0	800	64	80	64	0.42	0.75
RF-GlutarySite	338	0.001	800	69	70	68	0.38	0.77
RF-GlutarySite	128	0.002	800	72	73	70	0.43	0.81
RF-GlutarySite	128	0.003	800	72	73	70	0.43	0.81
RF-GlutarySite	69	0.004	800	71	75	68	0.42	0.79

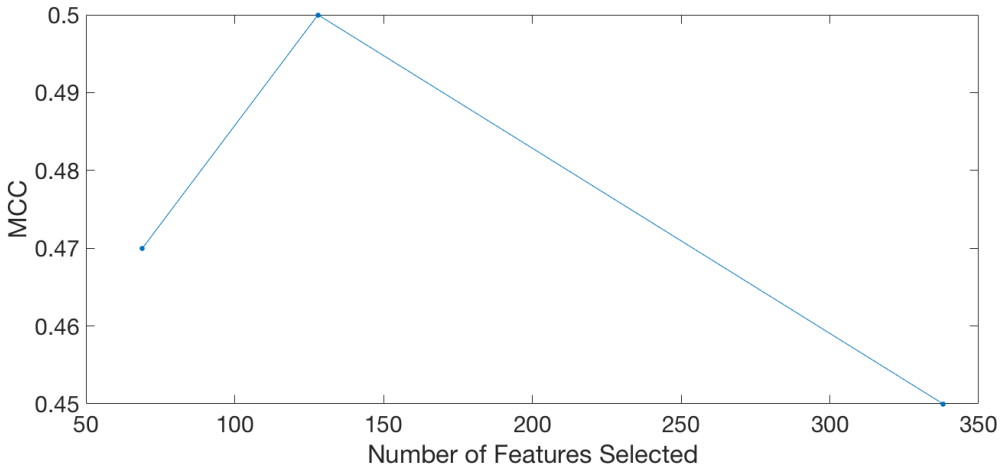

***Fig. S9**. Shows different number of feature selected against MCC metrics based on 10-fold cross validation.*

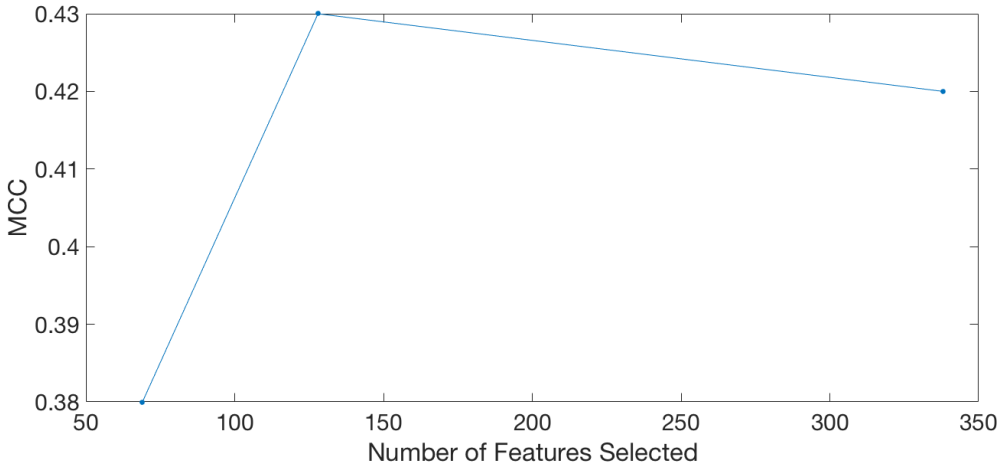

***Fig. S10**. Shows different number of feature selected against MCC metrics based on an independent set.*

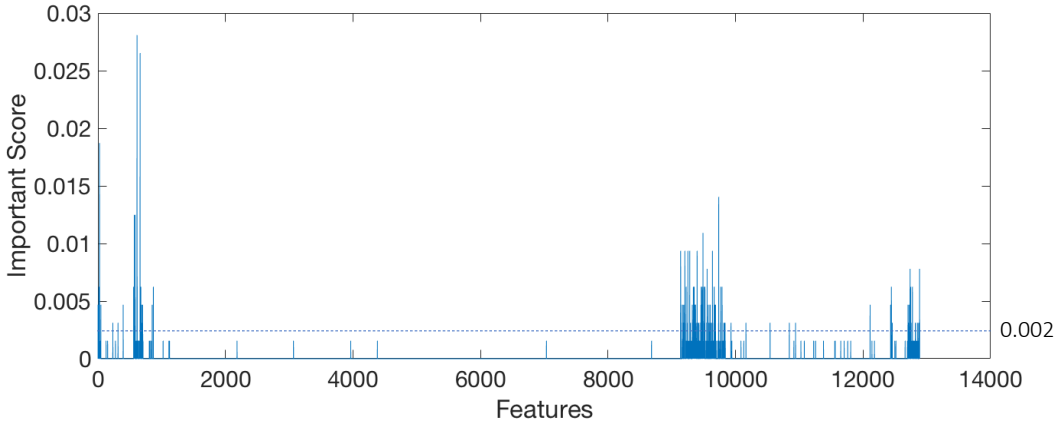

Fig. S11*. Shows highly correlated features with glutarylation sites. A threshold of 0.002 (dashed line) was chosen based on the comparative analysis outlined above.*

References

1. Vacic, Vladimir, Lilia M. Iakoucheva, and Predrag Radivojac. "Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments." *Bioinformatics* 22.12 (2006): 1536-1537.