Received August 1, 2020, accepted August 15, 2020, date of publication September 8, 2020, date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022629

N-GlycoGo: Predicting Protein N-Glycosylation Sites on Imbalanced Data Sets by Using Heterogeneous and Comprehensive Strategy

CHING-HSUAN CHIEN[1,2], CHI-CHANG CHANG[3,4], SHIH-HUAN LIN[2,5], CHI-WEI CHEN[1,6], ZONG-HAN CHANG[1], AND YEN-WEI CHU[1,2,7,8,9,10,11]

[1] Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan
[2] Ph.D. Program in Medical Biotechnology, National Chung Hsing University, Taichung 402, Taiwan
[3] School of Medical Informatics, Chung Shan Medical University, Taichung 402, Taiwan
[4] IT Office, Chung Shan Medical University Hospital, Taichung 402, Taiwan
[5] Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan
[6] Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan
[7] Institute of Molecular Biology, National Chung Hsing University, Taichung 402, Taiwan
[8] Agricultural Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan
[9] Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan
[10] Ph.D. Program in Translational Medicine, National Chung Hsing University, Taichung 402, Taiwan
[11] Rong Hsing Research Center for Translational Medicine, Taichung 402, Taiwan

Corresponding author: Yen-Wei Chu (ywchu@nchu.edu.tw)

This research was supported by the Ministry of Science and Technology, Taiwan, R.O.C. [grant numbers 108-2321-B-005-008, 108-2634-F-005-002], and the National Chung Hsing University and Chung-Shan Medical University [grant number NCHU-CSMU-10811].

ABSTRACT Glycosylation is the most complex post-modification effect of proteins. It participates in many biological processes in the human body and is closely related to many disease states. Among them, N-linked glycosylation is the most contained glycosylation data. However, the current N-linked glycosylation prediction tool does not take into account the serious imbalance between positive and negative data. In this study, we used protein sequence and amino acid characteristics to construct an N-linked glycosylation prediction model called N-GlycoGo. Based on sequence, structure, and function, 11 heterogeneous features were encoded. Further, XGBoost was selected for modeling. Finally, independent testing of human and mouse prediction models showed that N-GlycoGo is superior to other tools with Matthews correlation coefficient (MCC) values of 0.397 and 0.719, respectively, which is higher than other glycosylation site prediction tools. We have developed a fast and accurate prediction tool, N-GlycoGo, for N-linked glycosylation. N-GlycoGo is available at http://ncblab.nchu.edu.tw/n-glycogo/.

INDEX TERMS Ensemble learning, machine learning, N-linked glycosylation.

I. INTRODUCTION

Glycosylation is the most complex and common post-translational modification and involves the enzymatic attachment of sugars to proteins. Glycosylation affects many important biological processes like protein folding, cell-to-cell information transmission, gene expression, and control of cellular metabolism. Four main types of glycosylation patterns are known: N-linked, O-linked, C-linked, and GPI anchors. N-linked glycosylation, the most common, involves the attachment of carbohydrates to the amine group (NH2) of asparagine at the conserved motifs N-X-S and N-X-T, where X can be any amino acid except proline [1], [2]. To control and predict glycosylation, various genetic or cell culture methods of modification [3] and dynamics [4], genetic engineering [5], and genome models [6] have been used. the construction of these models requires computing tools and biological experimental methods and parameter adjustment training and repeated experiments require considerable time, especially for the mechanistic kinetic models [7]. Although these technologies have high accuracy, the instrument is expensive. Moreover, the large amount of data generated consumes considerable experimental material and labor. Therefore, using machine learning methods to develop tools for predicting glycosylation sites within a few hours is essential. Several prediction tools use amino acid sequences to predict post-translational modification sites.

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

Publicly available glycosylation prediction tools include NetNGlyc [8], GPP [9], GlycoPP [10], GlycoEP [11], SPRINT-Gly [12], and N-GlyDE [13]. NetNGlyc 1.0 uses artificial neural networks (ANNs) to predict the N-glycosylation sites on human proteins. GPP employs secondary structure (SS) and surface accessibility (ASA) [14] of mammalian protein sequences and then uses random forest (RF) prediction. GlycoPP performs binary profile of patterns (BPP), composition profile of patterns (CPP), and PSSM profile of patterns (PPP) for human protein sequences and then uses support vector machine (SVM) for prediction. GlycoEP performs BPP, CPP, PPP, SS, and ASA coding for eukaryotic protein sequences and then uses SVM to predict. SPRINT-Gly uses deep neural networks (DNNs) to predict glycosylation sites on N-linked and O-linked human and mouse protein sequences. N-GlyDE uses SVM to generate a two-stage prediction model for human glycosylation. Although the current prediction methods are accurate, some problems remain, such as the dataset of the training model is relatively small, amino acid information used for feature encoding is incomplete, and feature selection technology is not used to remove unimportant feature values. The choice of classifier also uses older methods; and several new classification algorithms are available that can greatly improve accuracy.

Therefore, we constructed N-GlycoGo to improve the prediction method of glycosylation sites through integrated models [15], use all positive and negative imbalance data, solve the problem of imbalanced data, and develop a more accurate model for predicting glycosylation. In addition to the sequence and structure based features for feature encoding, the subcellular location of a protein contains important information about protein function and is closely related to the signal peptide [16]. Therefore, SignalP-5.0 [17] has been added as a function-based feature. N-GlycoGo uses a total of five coding tools to generate 11 features. XGBOOST [18] was used to build a prediction model for N-linked glycosylation sites. In the independent tests of humans and mice, the highest MCC was 0.957 and 0.738, respectively. From the performance of other tools, it can be seen that the early tools have lower MCC, and glycosylation sites cannot be predicted across species.

II. MATERIALS AND METHODS

N-GlycoGo uses an ensemble model [19] to predict using heterogeneous features.. The flowchart for constructing N-linked glycosylation prediction tools for humans and mice is shown in Figure 1.

A. DATA COLLECTION

The glycosylation data sources used by N-GlycoGo include Universal Protein Resource (UniProt), dbPTM, and O-GlycBase v6.00.

1) UNIPROT

UniProt [20] is a database of protein sequence and annotation data jointly developed by the European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and Protein Information Resources (PIR), which includes UniProtKB [21], UniRef [22], and UniParc [23].

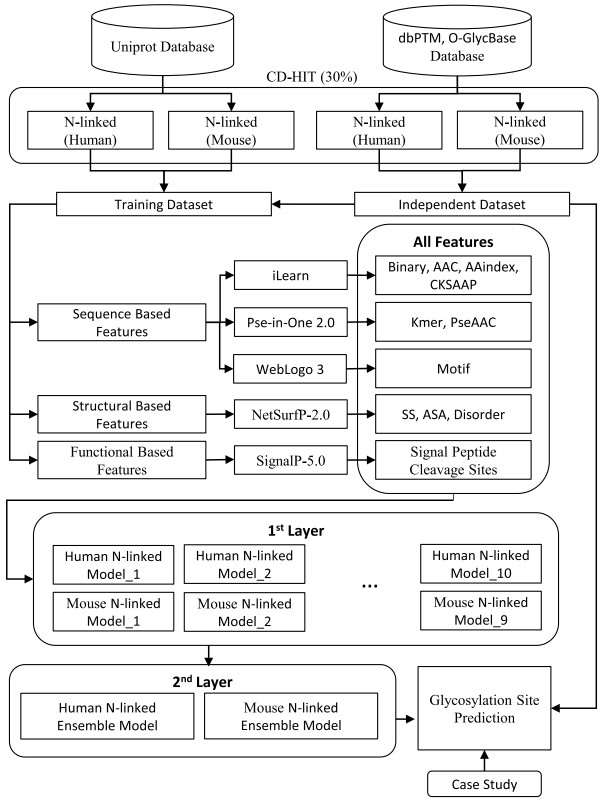

FIGURE 1. Flow chart of N-GlycoGo.

2) DBPTM

dbPTM [24] is a post-translational modifications database that integrates experimentally verified data from multiple databases.

3) O-GLYCBASE V6.00

O-GlycBase v6.00 [25] is a non-redundant glycosylation database and contains 242 protein sequences.

B. DATA PREPARATION

1) HUMAN TRAINING SET

The training set used by N-GlycoGo was obtained from UniProt. Using the post-translational modification information database, search with the keyword glycosylation was done and data labelled CARBOHYD and verified by experiments was annotated (excluding the annotation lines labeled probable, potential, and similar). The experimentally verified glycosylated or non-glycosylated N-linked sites were considered positive and negative sites, respectively. The sequences were fragmented with 21 window size and the glycosylation action site was placed at the center with 10 amino acids on the left and right each. for a blank value, a virtual amino acid "-" was added. Thereafter, CD-HIT was used to remove sequences that were more than 30% similar to avoid machine learning over-evaluation. A total of 3836 positive and 18277 negative sites were obtained for humans.

2) MOUSE TRAINING SET

The same is protocol was used to obtain mouse protein data from UniProt as that used for the human training set. A total of 57 positive and 948 negative sites were obtained for mice.

3) HUMAN INDEPENDENT SET

Glycosylation data from different sources were used to evaluate the stability of the model. For humans, data was collected from dbPTM and O-GlycBase. Next, after removing the proteins that appeared in the human training set, CD-HIT was used to remove more than 30% similar sequences and a total of 57 glycosylation sites remained. Thereafter, positive and negative sites were extracted to yield 57 positive and 948 negative sites.

4) MOUSE INDEPENDENT SET

Glycosylation sites of mouse protein data from dbPTM and O-GlycBase were selected for evaluation. Next, after removing the proteins that appeared in the mouse training set, CD-HIT was used to remove more than 30% similar sequences. Finally, 13 glycosylation sites, including 13 positive sites and 145 negative sites were selected.

C. PREDICTIVE MODEL

In the training and independent testing data, the difference in the ratio between positives and negatives is clear. Therefore, ensemble learning is used to construct the model to solve the problem of imbalanced data [19]. N-GlycoGo uses ensemble learning to extract samples from negatives so that the number of negatives and positives for each model are similar; finally, these models are integrated to improve the overall performance. The constructed algorithm includes Random Forest, SVM, and XGBoost.

D. FEATURE ENCODING

N-GlycoGo uses five coding tools to generate 11 features and is divided into three categories: sequence-, structure-, and function-based features.

1) SEQUENCE-BASED FEATURES

iLearn [26] can encode through DNA, RNA, and protein sequences. We used iLearn's binary, AAindex [27], amino acid composition (AAC) [28] and the composition of k-spaced amino acid pairs (CKSAAP) [29]. Binary encodes amino acids in a binary manner. The 20 amino acids are converted into 0 and 1 with 20-dimensional vector encoding to form 20 different combinations of sequence codes; window size 21 is used for sequence encoding. Features of 420 bits are used. It can be the most primitive and direct expression of the composition and distribution of the linear amino acid sequence. AAindex is a database for the physical and biochemical properties of amino acids. It is divided into three sections: AAindex1, AAindex2, and AAindex3. N-GlycoGo only uses AAindex1 because glycosylation is related to peptide binding and has nothing to do with amino acid mutations (AAindex2). Moreover, these peptides are linear and do not form secondary structures (AAindex3). Therefore, only AAindex1 is used. The 531 physical, chemical, and biochemical properties of the data are coded as features. The AAC code calculates the frequency of each amino acid type in a protein or peptide sequence. CKSAAP coding calculates the frequency of amino acid pairs separated by k residues ($k = 0, 1, 2, \ldots, 5$. The default maximum value of k is 5).

The Pse-in-One [30] tool was developed by the Harbin Institute of Technology, and can generate pseudo components of DNA, RNA, and protein sequences. We used three protein prediction modules of this tool—Kmer, parallel correlation pseudo amino acid composition (PC-PseAAC), series correlation pseudo amino acid composition (SC-PseAAC)—and made evaluations according to the output results, taking into account complete protein sequence and window size 21 sequence features to increase feature information. The value of Kmer represents the occurrence frequencies of k adjacent amino acids. PC-PseAAC combines continuous local sequence-order information and global sequence-order information into protein sequence feature vectors. SC-PseAAC is a variant of PC-PseAAC, which combines local sequence-order information and global sequence-order information into a protein sequence feature vector.

WebLogo 3 [31] displays multiple sequences of amino acids or nucleic acids through alignments. The amino acid at each position in the sequence can be stacked with the English abbreviation of the nucleic acid, and the height of the stacked letters represents the relative frequency of the amino acid or nucleic acid at that position. The glycosylation site is conserved. Previous studies on N-linked glycosylation have reported that the glycosylation site N-X-S or N-X-T is conserved, where X can be any amino acid except proline. We used WebLogo 3 to evaluate the frequency value of each amino acid in the positive segment. For a gap, the value was 0.

2) STRUCTURE-BASED FEATURES

NetsurfP-2.0 [32] can predict the structural characteristics of the protein or amino acid sequences through deep learning, including the surface accessibility data for exposed and embedded amino acids, probability of α-helix, β-strand and random coil, data for structural disorder of proteins [33], and phi/psi value of dihedral angles [34] for amino acids.

Protein surface accessibility (relative/absolute surface accessibility, RSA/ASA) includes evaluation of buried or exposed residues and ASA Z-score (Z-score is a prediction of surface area and does not contain structural information). Buried and exposed residues are scored as 10 and 01, respectively, and the Z-scores of RSA/ASA and ASA are added to the score. The secondary structure provides the possibility scores for α-helix, β-strand, and random coil, and the three values are used for scoring.

3) FUNCTIONAL-BASED FEATURES

SignalP 5.0 [17] is based on the amino acid sequence of archaea, gram-positive bacteria, gram-negative bacteria, and eukaryotic proteins through a deep neural network to predict signal peptides (SP) [35] cleavage site. The subcellular localization of the protein depends on the signal peptide [16]. SignalP 5.0 is used to predict the signal peptide cleavage site on the sequence. The prediction result includes the C-score, the score of the original cleavage site, and the S-score. The signal peptide score and Y-score include the score of the cleavage site. Moreover, three values are evaluated.

E. FEATURE SELECTION

N-GlycoGo uses mRMR for feature selection. mRMR is a feature filter, where ''relevance'' and ''redundancy'' are defined using mutual information, correlation, t-test/F-test, distance, etc. A total of three feature ranking results are the output, including max-relevance and MRMR calculated using two schemes of mutual information difference (miD) and mutual information quotient (miQ).

F. ALGORITHMIC ENSEMBLE TECHNIQUES

The simple integration method continuously draws samples from the majority class, making the number of samples of the majority and minority classes the same and finally integrates these models. The main purpose of the ensemble method is to improve the performance of a single classifier. This method constructs several two-level classifiers from the original data and then assembles the predicted results.

G. MODEL EVALUATION

To judge the quality of the model requires certain criteria; therefore, the choice of evaluation indicators is also important. Accuracy (ACC), sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC) are common indicators used to evaluate machine learning. ACC is the most intuitive evaluation indicator when evaluating models, as shown in equation (1), where TP, FP, FN, and TN, are true positives, false positives, false negatives, and true negatives, respectively. Sn represents the proportion of all positives that are correctly predicted, as shown in equation (2), which reflects the model's ability to predict positives. Sp represents the ratio of all correctly predicted negatives. As shown in equation (3), this ratio shows the model's ability to correctly predict negatives. MCC is a suitable evaluation index when the ratio of positives to negatives is not even. The value of MCC approaches 0, when almost all the predictions are wrong; MCC is equal to 1, when all predictions are correct; MCC $= -1$, when all predicted results and actual values are opposite, as shown in formula (4).

$$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{3}$$

$$MCC = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{4}$$

III. RESULTS AND DISCUSSION

A. COMPARISON OF ALGORITHM

To solve the problem of data imbalance, we constructed a model in ensemble learning. The prediction method was evaluated by ten-fold cross-validation, as shown in Table 1. For data with CD-HIT deduplication, XGBoost can reach 0.981 in MCC, which is much higher than 0.96 of traditional SVM and 0.961 of RandomForest.

TABLE 1. Comparison with traditional methods after adding features.

Algorithm	MCC
SVM	0.96
RandomForest	0.961
XGBoost	0.981

B. FEATURE ANALYSIS

To explore the importance of each feature in predicting N-linked glycosylation sites in humans, N-GlycoGo uses mRMR to test the top 10 and 100 features from 14383 features for modeling and prediction, and accuracy calculation.

TABLE 2. Feature selection and model stability.

Top N	Layer	Assessment	Mcc		
			MaxRel	MID	MIQ
N = 10	First layer model	Average Value	0.397	0.392	0.389
		Standard Deviation	0.000	0.003	0.001
	Second layer model	Average Value	0.397	0.394	0.389
		Standard Deviation	0.000	0.001	0.001
N = 100	First layer model	Average Value	0.397	0.395	0.392
		Standard Deviation	0.000	0.000	0.005
	Second layer model	Average Value	0.397	0.395	0.390
		Standard Deviation	0.000	0.000	0.007

As seen in Table 2, the ensemble learning construction model can slightly improve MCC when there are a small number of features, only when the features selected by MIQ schemes are selected, the MCC decreases; however, when the number of features reaches 100, because the features are taken by other schemes, the ensemble learning models have the same and stable predictions; therefore, second stage predictions cannot improve the accuracy of the predictions.

As seen in Table 2, selection of more features does not necessarily increase accuracy and slows down the running speed. Choosing the right number of features can increase the prediction speed and accuracy of the data.

TABLE 3. Forecast site comparison.

Tools	Years	Type of glycosylation	Classifier	Species
NetNGlyc	2002	N-linked	Neural networks	Human
GPP	2008	N-, O-linked	RF	Eukaryotic
GlycoPP	2012	N-, O-linked	SVM	Prokaryotic
GlycoEP	2013	N-, O-, C-linked	SVM	Eukaryotic
SPRINT-Gly	2019	N-, O-linked	Neural networks	Human and mouse
N-GlyDE	2019	N-linked	SVM	Human

TABLE 4. Comparison of predictors with independent test of humans.

Human	TP	FP	TN	FN	Sn	Sp	Acc	MCC
NetNGlyc	43	746	227	15	0.741	0.233	0.262	-0.014
GPP	51	386	587	7	0.879	0.603	0.619	0.225
GlycoPP Avg.	31.8	347.3	625.8	26.3	0.548	0.643	0.638	0.098
GlycoPP_BPP	40	214	759	18	0.690	0.780	0.775	0.251
GlycoPP_CPP	25	448	525	33	0.431	0.540	0.533	-0.014
GlycoPP_PPP	7	119	854	51	0.121	0.878	0.835	-0.001
GlycoPP_ASA_BPP	55	608	365	3	0.948	0.375	0.407	0.156
GlycoEP Avg.	33.2	227.7	745.3	24.8	0.572	0.766	0.755	0.195
GlycoEP_S_BPP	48	149	824	10	0.828	0.847	0.846	0.395
GlycoEP_S_CPP	33	544	429	25	0.569	0.441	0.448	0.005
GlycoEP_S_PPP	42	257	716	16	0.724	0.736	0.735	0.234
GlycoEP_S_ASA_BPP	14	134	839	44	0.241	0.862	0.827	0.068
GlycoEP_A_BPP	48	148	825	10	0.828	0.848	0.847	0.397
GlycoEP_A_ASA_BPP	14	134	839	44	0.241	0.862	0.827	0.068
Sprint-Gly	46	147	826	12	0.793	0.849	0.846	0.379
N-GlyDE	31	73	900	27	0.534	0.925	0.903	0.352
N-GlycoGo	**48**	**148**	**825**	**10**	**0.828**	**0.848**	**0.847**	**0.397**

TABLE 5. Comparison of predictor with independent test of mouse.

Mouse	TP	FP	TN	FN	Sn	Sp	Acc	MCC
NetNGlyc	10	116	37	3	0.769	0.242	0.283	0.007
GPP	13	54	99	0	1.000	0.647	0.675	0.354
GlycoPP Avg.	8.3	62.0	91.0	4.8	0.635	0.595	0.598	0.142
GlycoPP_BPP	9	38	115	4	0.692	0.752	0.747	0.265
GlycoPP_CPP	9	97	56	4	0.692	0.366	0.392	0.033
GlycoPP_PPP	3	16	137	10	0.231	0.895	0.843	0.106
GlycoPP_ASA_BPP	12	97	56	1	0.923	0.366	0.410	0.164
GlycoEP Avg.	9.3	30.3	122.7	3.7	0.718	0.802	0.795	0.382
GlycoEP_S_BPP	13	8	145	0	1.000	0.948	0.952	0.766
GlycoEP_S_CPP	12	93	60	1	0.923	0.392	0.434	0.176
GlycoEP_S_PPP	12	34	119	1	0.923	0.778	0.789	0.421
GlycoEP_S_ASA_BPP	3	19	134	10	0.231	0.876	0.825	0.084
GlycoEP_A_BPP	13	8	145	0	1.000	0.948	0.952	0.766
GlycoEP_A_ASA_BPP	3	20	133	10	0.231	0.869	0.819	0.078
Sprint-Gly	12	10	143	1	0.923	0.935	0.934	0.680
N-GlyDE	9	4	149	4	0.692	0.974	0.952	0.666
N-GlycoGo	**12**	**8**	**145**	**1**	**0.923**	**0.948**	**0.946**	**0.719**

C. PERFORMANCE OF INDEPENDENT TEST

We have compiled the existing prediction models for glycosylation sites in Table 3. The table contains the modeling methods, type of glycosylation, and species for each model. NetNGlyc [8], GPP [9], GlycoPP [10], GlycoEP [11], SPRINT-Gly [12], and N-GlyDE [13] were selected for this study.

NetNGlyc 1.0 uses artificial neural networks (ANN) to predict the glycosylation sites on N-linked human protein sequences. However, it can only predict data for which sequence length is less than 2000. GPP uses SS and ASA to score mammalian protein sequences and uses random forest to predict. GlycoPP uses BPP, CPP, PPP, and ASA + BPP to predict glycosylation sites in prokaryotes through SVM. GlycoEP uses features such as BPP, CPP, PPP, and ASA + BPP to predict through SVM, and provides four features for users to choose. According to the training set, it is divided into two prediction tools: Standard Predictor (S) and Advanced Predictor (A). SPRINT-Gly uses a Deep neural network (DNN) to predict glycosylation sites on N-linked and O-linked human and mouse protein sequences. N-GlyDE uses SVM to carry out a two-stage sequence prediction model. The first stage provides a prediction score for each protein, and the second stage glycosylation prediction score can be adjusted according to the prediction score.

To evaluate the predictive performance and stability of N-GlycoGo, the protein sequence in the independent set was used for prediction, Sn, Sp, ACC, and MCC were calculated based on the prediction results, and the existing glycosylation site predictions were used for comparison. The accuracy of N-GlycoGo for the independent set in human is shown in Table 4. The MCC value is 0.397, which is tied for first place with GlycoEP_A_BPP. The accuracy of the independent set for mouse is shown in Table 5. GlycoEP's BPP has the highest MCC value of 0.766, followed by N-GlycoGo's 0.719. But the performance of GlycoEP in 6 different prediction models with large variation. The average MCC of GlycoEP is only 0.382. It may be difficult for users to choose a suitable prediction model. NetNGlyc is the earliest glycosylation prediction tool. The early data is relatively incomplete and so the MCC is the lowest. GPP is also an early prediction tool. GlycoPP targets prokaryotes and does not perform well for eukaryotes, such as humans and mice. The glycosylation sites of different species are different. GlycoEP provides multiple prediction models and the differences between the models are very large. Sprint-Gly establishes prediction models for mice and humans, whereas N-GlyDE establishes prediction models for humans. Sprint-Gly and N-GlyDE were released in 2019 and have better performance than other tools.

IV. CONCLUSION

N-GlycoGo is based on the ensemble learning model. It uses information from human and mouse N-linked glycosylation sites and considers sequence based features, structure based features, and function based features. A total of 11 feature codes are present. The best model is integrated with the relevant information. First, Binary, AAindex, AAC, CKSAAP, Kmer, PC-PseAAC, SC-PseAAC, Motif, RSA/ASA, SS, and SignalP are encoded by 21 window size amino acid fragments; the results are predicted using various integrated models through tenfold cross-validation and XGBoost performed best with an MCC of 0.968. Using the independent set evaluation model compiled by dbPTM and O-GlycBase, which is different from the training data set, XGBoost can also reach an MCC of 0.397 and 0.719 in human and mouse, respectively. Therefore, XGBoost is used as the basic model for N-GlycoGo prediction.

The independent set was used for existing glycosylation site prediction websites, including NetNGlyc, GPP, GlycoEP, GlycoPP, SPRINT-Gly, and N-GlyDE. For accuracy evaluation conducted using the independent set for human. The MCC values of N-GlycoGo and GlycoEP_A_BPP are tied for first place. For accuracy evaluation, conducted by using the independent set of mouse, all performance of other tools were lower than that ofN-GlycoGo except GlycoEP's BPP. N-GlycoGo was much higher than the average MCC of different models of GlycoEP.

N-GlycoGo was developed by strictly analyzing and integrating the best architecture in each step for glycosylation site prediction in human and mouse. It will help researchers to reduce time greatly and predict accurately.

REFERENCES

[1] Y. Gavel and G. Von Heijne, "Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: Implications for protein engineering," *Protein Eng., Des. Selection*, vol. 3, no. 5, pp. 433–442, 1990, doi: 10.1093/protein/3.5.433.

[2] M. Kowarik, N. M. Young, S. Numao, B. L. Schulz, I. Hug, N. Callewaert, D. C. Mills, D. C. Watson, M. Hernandez, J. F. Kelly, M. Wacker, and M. Aebi, "Definition of the bacterial N-glycosylation site consensus sequence," *EMBO J.*, vol. 25, no. 9, pp. 1957–1966, May 2006, doi: 10.1038/sj.emboj.7601087.

[3] P. Hossler, "Protein glycosylation control in mammalian cell culture: Past precedents and contemporary prospects," in *Genomics and Systems Biology of Mammalian Cell Culture*, W. S. Hu and A.-P. Zeng, Eds. Berlin, Germany: Springer, 2012, pp. 187–219.

[4] I. J. del Val, J. M. Nagy, and C. Kontoravdi, "A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing golgi apparatus," *Biotechnol. Prog.*, vol. 27, no. 6, pp. 1730–1743, Nov. 2011, doi: 10.1002/btpr.688.

[5] A. G. McDonald, J. M. Hayes, T. Bezak, S. A. G Uchowska, E. F. J. Cosgrave, W. B. Struwe, C. J. M. Stroop, H. Kok, T. van de Laar, P. M. Rudd, K. F. Tipton, and G. P. Davey, "Galactosyltransferase 4 is a major control point for glycan branching in N-linked glycosylation," *J. Cell Sci.*, vol. 127, no. 23, pp. 5014–5026, Dec. 2014, doi: 10.1242/jcs.151878.

[6] B. G. Kremkow and K. H. Lee, "Glyco-mapper: A Chinese hamster ovary (CHO) genome-specific glycosylation prediction tool," *Metabolic Eng.*, vol. 47, pp. 134–142, May 2018, doi: 10.1016/j.ymben.2018.03.002.

[7] G. L. Medlock and J. A. Papin, "Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning," *Cell Syst.*, vol. 10, no. 1, pp. 109.e3–119.e3, Jan. 2020, doi: 10.1016/j.cels.2019.11.006.

[8] R. Gupta, E. Jung, and S. Brunak. (2004). *NetNGlyc 1.0 Server: Prediction of N-Glycosylation Sites in Human Proteins*. Accessed: Oct. 30, 2015. [Online]. Available: http://www.cbs.dtu.dk/services/NetNGlyc/

[9] S. E. Hamby and J. D. Hirst, "Prediction of glycosylation sites using random forests," *BMC Bioinf.*, vol. 9, no. 1, p. 500, Nov. 2008, doi: 10.1186/1471-2105-9-500.

[10] J. S. Chauhan, A. H. Bhat, G. P. S. Raghava, and A. Rao, "GlycoPP: A webserver for prediction of N- and O-glycosites in prokaryotic protein sequences," *PLoS ONE*, vol. 7, no. 7, Jul. 2012, Art. no. e40155, doi: 10.1371/journal.pone.0040155.

[11] J. S. Chauhan, A. Rao, and G. P. S. Raghava, "In silico platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e67008, doi: 10.1371/journal.pone.0067008.

[12] G. Taherzadeh, A. Dehzangi, M. Golchin, Y. Zhou, and M. P. Campbell, "SPRINT-gly: Predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties," *Bioinformatics*, vol. 35, no. 20, pp. 4140–4146, Oct. 2019, doi: 10.1093/bioinformatics/btz215.

[13] T. Pitti, C.-T. Chen, H.-N. Lin, W.-K. Choong, W.-L. Hsu, and T.-Y. Sung, "N-GlyDE: A two-stage N-linked glycosylation site prediction incorporating gapped dipeptides and pattern-based encoding," *Sci. Rep.*, vol. 9, no. 1, Nov. 2019, Art. no. 15975, doi: 10.1038/s41598-019-52341-z.

[14] H. Naderi-Manesh, M. Sadeghi, S. Arab, and A. A. M. Movahedi, "Prediction of protein surface accessibility with information theory," *Proteins, Struct., Function, Bioinf.*, vol. 42, no. 4, pp. 452–459, 2001, doi: 10.1002/1097-0134(20010301)42:4<452::AID-PROT40>3.0.CO;2-Q.

[15] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar. 2009, pp. 324–331, doi: 10.1109/CIDM.2009.4938667.

[16] D. Nguyen, R. Stutz, S. Schorr, S. Lang, S. Pfeffer, H. H. Freeze, F. Förster, V. Helms, J. Dudek, and R. Zimmermann, "Proteomics reveals signal peptide features determining the client specificity in human TRAP-dependent ER protein import," *Nature Commun.*, vol. 9, no. 1, Sep. 2018, Art. no. 3765, doi: 10.1038/s41467-018-06188-z.

[17] J. J. A. Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 5.0 improves signal peptide predictions using deep neural networks," *Nature Biotechnol.*, vol. 37, no. 4, pp. 420–423, Apr. 2019, doi: 10.1038/s41587-019-0036-z.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[19] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, Boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012, doi: 10.1109/TSMCC.2011.2161285.

[20] The UniProt Consortium, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, Jan. 2017, doi: 10.1093/nar/gkw1099.

[21] E. Boutet *et al.*, "UniProtKB/swiss-prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view," *Methods Mol. Biol. Clifton NJ*, vol. 1374, pp. 23–54, 2016, doi: 10.1007/978-1-4939-3167-5_2.

[22] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: Comprehensive and non-redundant UniProt reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282–1288, May 2007, doi: 10.1093/bioinformatics/btm098.

[23] R. Leinonen, F. G. Diez, D. Binns, W. Fleischmann, R. Lopez, and R. Apweiler, "UniProt archive," *Bioinformatics*, vol. 20, no. 17, pp. 3236–3237, Nov. 2004, doi: 10.1093/bioinformatics/bth191.

[24] K.-Y. Huang, T.-Y. Lee, H.-J. Kao, C.-T. Ma, C.-C. Lee, T.-H. Lin, W.-C. Chang, and H.-D. Huang, "DbPTM in 2019: Exploring disease association and cross-talk of post-translational modifications," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D298–D308, Jan. 2019, doi: 10.1093/nar/gky1074.

[25] R. Gupta, H. Birch, K. Rapacki, S. Brunak, and J. E. Hansen, ''O-GLYCBASE version 4.0: A revised database of O-glycosylated proteins,'' *Nucleic Acids Res.*, vol. 27, no. 1, pp. 370–372, Jan. 1999, doi: 10.1093/nar/27.1.370.

[26] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, ''iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data,'' *Briefings Bioinf.*, vol. 21, no. 3, pp. 1047–1057, May 2020, doi: 10.1093/bib/bbz041.

[27] S. Kawashima, H. Ogata, and M. Kanehisa, ''AAindex: Amino acid index database,'' *Nucleic Acids Res.*, vol. 27, no. 1, pp. 368–369, Jan. 1999.

[28] K.-C. Chou and C.-T. Zhang, ''A correlation-coefficient method to predicting protein-structural classes from amino acid compositions,'' *Eur. J. Biochem.*, vol. 207, no. 2, pp. 429–433, Jul. 1992, doi: 10.1111/j.1432-1033.1992.tb17067.x.

[29] Y.-Z. Chen, Y.-R. Tang, Z.-Y. Sheng, and Z. Zhang, ''Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs,'' *BMC Bioinf.*, vol. 9, no. 1, p. 101, 2008, doi: 10.1186/1471-2105-9-101.

[30] B. Liu, H. Wu, and K.-C. Chou, ''Pse-in-one 2.0: An improved package of Web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences,'' *Natural Sci.*, vol. 9, no. 4, pp. 67–91, 2017, doi: 10.4236/ns.2017.94007.

[31] G. E. Crooks, ''WebLogo: A sequence logo generator,'' *Genome Res.*, vol. 14, no. 6, pp. 1188–1190, May 2004, doi: 10.1101/gr.849004.

[32] M. S. Klausen, M. C. Jespersen, H. Nielsen, K. K. Jensen, V. I. Jurtz, C. K. Sønderby, M. O. A. Sommer, O. Winther, M. Nielsen, B. Petersen, and P. Marcatili, ''NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning,'' *Proteins, Struct., Function, Bioinf.*, vol. 87, no. 6, pp. 520–527, Jun. 2019, doi: 10.1002/prot.25674.

[33] R. Pancsa and P. Tompa, ''Structural disorder in eukaryotes,'' *PLoS ONE*, vol. 7, no. 4, Apr. 2012, Art. no. e34687, doi: 10.1371/journal.pone.0034687.

[34] M. J. Wood and J. D. Hirst, ''Protein secondary structure prediction with dihedral angles,'' *Proteins, Struct., Function, Bioinf.*, vol. 59, no. 3, pp. 476–481, Mar. 2005, doi: 10.1002/prot.20435.

[35] J. W. Izard and D. A. Kendall, ''Signal peptides: Exquisitely designed transport promoters,'' *Mol. Microbiol.*, vol. 13, no. 5, pp. 765–773, Sep. 1994, doi: 10.1111/j.1365-2958.1994.tb00469.x.

CHING-HSUAN CHIEN received the M.S. degree in institute of genomics and bioinformatics from National Chung Hsing University, Taichung, Taiwan, in 2018, where she is currently pursuing the Ph.D. degree in Ph.D. program in medical biotechnology. Her research area is bioinformatic.

CHI-CHANG CHANG is currently Professor with the Medical Informatics School, Chung Shan Medical University, also consultant on Smart Healthcare Committee and Cancer Center, Chung Shan Medical University Hospital. He has published more than 100 journal articles in reputed international journals. His research interests include second primary cancers and recurrent cancers, bioinformatics, digital traditional Chinese medicine, and clinical operational research.

Prof. Chang has served as the Guest Editor for many special issues include, the *International Journal of Environmental Research and Public Health* (SCI, IF 2.849), the *International Journal of Medical Sciences* (SCI, IF 2.523), *Therapeutics and Clinical Risk Management* (SCI, IF 1.888), the *Journal of Universal Computer Science* (SCI, IF 0.701), and *Open Medicine* (SCI, IF 1.204). He is the Founding Chair of the International Conference on Medical and Health Informatics (ICMHI), the International Conference on Healthcare Service Management (ICHSM), the Joint Executive Board on Medical Informatics in Taiwan (JEBMI), and Health Big Data, Analytics and Artificial Intelligence Forum (HBAAI).

SHIH-HUAN LIN received the M.S. degree in applied mathematics from National Chung Hsing University, Taiwan, where he is currently pursuing the Ph.D. degree in Ph.D. program in medical biotechnology. His research interests include algorithm development, artificial intelligence, numerical analysis, and preventive medicine.

CHI-WEI CHEN received the M.S. degree in bioinformatics from National Chung Hsing University, Taiwan, in 2012, where he is currently pursuing the Ph.D. degree in computer science. His research interests include bioinformatics and machine learning.

ZONG-HAN CHANG received the B.S. degree in biotechnology from Asia University, Taichung, Taiwan, in 2018, and the M.S. degree in institute of genomics and bioinformatics from National Chung Hsing University, Taichung, in 2020. His research areas are bioinformatics and machine learning.

YEN-WEI CHU received the Ph.D. degree in computer science from National Chiao Tung University, Taiwan, in 2006. His lab takes the technologies of data mining, machine learning, and artificial intelligence as the core algorithms. To establish a variety of intelligent decision-making systems for different issues, which is the important part for industry 4.0 and the Internet of Things. He is currently a Professor with the Institute of Genomics and Bioinformatics and the joint appointment Professor with the Institute of Molecular Biology, National Chung Hsing University, Taiwan. He mainly focus on building the learning model of humanoid intelligence, covering the field of bioinformatics, medical science, agriculture, food science, business, and astronomy. His research interests include bioinformatics algorithms, computational epigenetics, artificial intelligence, and intelligent systems.