Sequence analysis

Identification of sub-Golgi protein localization by use of deep representation learning features

Zhibin Lv 🟠 1,†, Pingping Wang 4,†, Quan Zou 🟢 1,2,3,* and Qinghua Jiang 4,*

1Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, 2Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, 3Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China and 4Center for Bioinformatics, School of Life Science and Technology, Harbin Institute of Technology, Harbin 150000, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Received on August 31, 2020; revised on December 10, 2020; editorial decision on December 13, 2020; accepted on December 14, 2020

Abstract

Motivation: The Golgi apparatus has a key functional role in protein biosynthesis within the eukaryotic cell with malfunction resulting in various neurodegenerative diseases. For a better understanding of the Golgi apparatus, it is essential to identification of sub-Golgi protein localization. Although some machine learning methods have been used to identify sub-Golgi localization proteins by sequence representation fusion, more accurate sub-Golgi protein identification is still challenging by existing methodology.

Results: we developed a protein sub-Golgi localization identification protocol using deep representation learning features with 107 dimensions. By this protocol, we demonstrated that instead of multi-type protein sequence feature representation fusion as in previous state-of-the-art sub-Golgi-protein localization classifiers, it is sufficient to exploit only one type of feature representation for more accurately identification of sub-Golgi proteins. Compared with independent testing results for benchmark datasets, our protocol is able to perform generally, reliably and robustly for sub-Golgi protein localization prediction.

Availability and implementation: A use-friendly webserver is freely accessible at http://isGP-DRLF.aibiochem.net and the prediction code is accessible at https://github.com/zhibinlv/isGP-DRLF.

Contact: zouquan@nclab.net or qhjiang@hit.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

As an important organelle in eukaryotic cells, the Golgi apparatus (GA) acts to process, sort and transport proteins synthesized by the endoplasmic reticulum (Tao et al., 2020), which are then delivered to specific compartments of the cell or secreted from the cell (Holthuis et al., 2014). GA malfunction can result in Parkinson’s disease (Fujita et al., 2006), Alzheimer’s disease (Gonatas et al., 1998) and other neurodegenerative disorders (Ligon et al., 2020). Thus it is essential to understand the functional details of the GA (Ravichandran et al., 2020) such as protein localization in the cis-Golgi (cis-Golgi protein) or in the trans-Golgi (trans-Golgi protein) (De Tito et al., 2020). Such an understanding would clarify GA function (Berry et al., 2017) and would provide clues to aid drug discovery and development (Stoeber et al., 2018).

Over the past decade, several machine learning-based methods for identification of sub-Golgi proteins localization (Yang et al., 2019b), including proteins cis- and trans-Golgi localization, have been developed using a few benchmark datasets (Ding et al., 2011, 2013; Yang et al., 2016b; Zhao et al., 2019). The previously reported protein sub-Golgi localization classifiers shared some common aspects to achieve high identification accuracy. The first was to use sequence feature fusion for better protein sequence representation (Ahmad et al., 2019; Rahman et al., 2018; Wang et al., 2020a). To the best of our knowledge, the most widely used features for high performance sub-Golgi protein classifiers are the fusion of a position specific scoring matrix (PSSM) (Jiao et al., 2016b; Shen et al., 2019a,b; Yang et al., 2016b), dipeptide composition frequency (Ding et al., 2011, 2013; Lv et al., 2019b; Rahman et al., 2018), pseudo amino acid physical and chemical properties (Jiao et al., 2016b,c; Zhao et al., 2019; Zhou et al., 2019) and their derivative features. The second was to use an over-sampling method to overcome the lack of sub-Golgi protein localization balance in training benchmark datasets (Ahmad et al., 2017, 2019; Lv et al., 2019b;
Rahman et al., 2018; Yang et al., 2016b). In addition, to feature fusion and data over-sampling, the employed selection technologies included analysis of variance (ANOVA) (Ding et al., 2016; Tang et al., 2018), Fisher method, minimum redundancy maximum relevance (mRMR), random forest recursive feature elimination (RF-RFE) and other methods that selected the best features in the vector space (Ahmad et al., 2017, 2019; Ding et al., 2013; Jiao et al., 2016c; Rahman et al., 2018; Wang et al., 2020b; Yang et al., 2016b). By combining the above techniques, the reported protein sub-Golgi localization classifier (isGPT) (Rahman et al., 2018) was with the best independent test score (ACC = 95.3%, MCC = 0.85, Sn = 84.6% and Sp = 98.0%). The good performance of isGPT was attained by carefully selected 2800 fusion features from the original feature space with 18 840 dimensions. The fusion features of isGPT were derived by combination of six types of protein sequence representation methods that included amino acid composition, dipeptides composition, tripeptide composition, n-gapped-dipeptides composition, position specific features and pseudo amino acid composition (Rahman et al., 2018).

Feature extraction plays an important role in protein sequence analysis and an appropriate and selected feature representation such as that used by isGPT (Rahman et al., 2018) greatly improves the accuracy of protein sequence analysis (Lv et al., 2019a). Given its automatic feature extraction and powerful feature representation capabilities, deep learning has been widely used in sequence analysis of proteins, DNA and RNA (Min et al., 2017; Wang et al., 2017; Xu, 2019; Xu et al., 2017, 2019). Deep learning is a form of machine learning that automatically learns feature representation by capturing parameters in a neural network (Eraslan et al., 2019; Jiang et al., 2019b). Based on this principle, pre-trained deep learning networks can be used for feature extraction from new data or migrated for application to other similar tasks such as image recognition and natural language processing, which are known as transfer learning (Zhang et al., 2016; Zhou et al., 2020). In 2019, Alley et al. (2019) proposed a self-supervised and universal protein sequence deep representation learning tool, UniRep, which was trained using UniRef50 (a dataset with tens of millions of protein sequences) to better represent natural and de-novo designed proteins. Also, some other preprint papers such as TAPE (Rao et al., 2019), BiLSTM embedding model (Bepler et al., 2019), PROBETa (Nambiar et al., 2020) and MULocDeep (Jiang et al., 2020a) have used similar ideas to encode protein sequences in a deep representations learning way and have obtained good results in many protein-sequence analysis applications.

In this work, we utilized UniRep to extract deep representation learning features for sub-Golgi protein sequences. Then, by using synthetic minority over-sampling (SMOTE) and light gradient boosting machine (LGBM) feature selection methods, we developed a new support vector machine based sub-Golgi protein localization classifier named as isGP-DRLF. The leave-one-out cross-validation scores of isGP-DRLF based on D5 dataset with only one type of feature representation vectors was ACC = 99.2%, MCC = 0.98, Sn = 100% and Sp = 98.4%, while the independent testing score metrics for isGP-DRLF was ACC = 96.4%, MCC = 0.90, Sp = 84.6% and Sn = 100%. The isGP-DRLF based on D3 dataset was with independent testing scores of ACC = 98.4%, MCC = 0.95, Sn = 100% and Sp = 98.0%, which improved by the relative value of 3.25%, 11.7%, 18.2% and 0.0% compared to the previously reported best independent-testing sub-Golgi protein classifier (isGPT) with six types of feature representation vectors. While iGPT used 2800-dimension fusion features for prediction, our isGP-DRLF used only 107-dimension features without any feature fusion. Herein, for sub-Golgi protein localization prediction, the effect of LGBM feature selection was better than that of ANOVA and MRMD feature selection technologies. The support vector machine algorithm is the best sub-Golgi protein identification algorithm in this study. A user-friendly isGP-DRLF webserver is available at http://isGP-DRLF.awbiochem.net for small sequence data. For a large dataset computing, the users could also download the trained model from https://github.com/zhibinlv/isGP-DRLF. By using UMAP(uniform manifold approximation and projection) method (McInnes et al., 2018) feature visualization technology, we found out that UniRep feature could be better to represent proteins than other features (single feature type or fused features types) to distinguish proteins in cis-Golgi from those in trans-Golgi. Thus, isGP-DRLF could employ merely one type of feature representation, a protein sequence deep representation learning feature, to carry out a powerful prediction tool for sub-Golgi proteins localization, although it abandoned the feature fusion methods as many state-of-the-art models did.

2 Materials and methods

2.1 Datasets

There are several benchmark datasets (Ding et al., 2011, 2013; Yang et al., 2016b; Zhao et al., 2019) with different sequence homologies and numbers for sub-Golgi protein identification modeling (see the details of datasets labeled D0, D1, D2, D3 and D4 in Supplementary Table S1). Considering the availability of the benchmark datasets and the performance of the sub-Golgi protein classifiers based on various datasets, we used benchmark dataset D3 for model training and D4 for model independent testing. D3 has been widely used for several state-of-the-art protein sub-Golgi localization classifiers (Ahmad et al., 2017, 2019; Rahman et al., 2018; Yang et al., 2016b), available at https://www.mdpi.com/1422-0067/17/22/4067 and D4 for independent testing available at http://lin-group.cn/server/SubGolgi/data. To prevent from homology or sequence similarity bias and from overfitting due to insufficient data entries, we have created a new updating training benchmark dataset downloaded from the latest Universal Protein KnowledgeBase (UniProtKB Version 2020.05) (Jiang et al., 2019a). We firstly searched protein sequence on the website https://www.uniprot.org/locations/ by using searching keywords listed in Supplementary Table S1 footnote a. Please see the dataset set up searching example in Supplementary Figure S0 in Supplementary Materials. Then we got a temporary dataset. The D4 was merged into temporary dataset. Then we removed the redundant sequences from the temporary dataset using PSI-CD-HIT (Jiang et al., 2020b) with a 25% identity cutoff. Finally, the independent testing sequences in D4 were excluded from the temporary dataset to avoid overfitting, and the new benchmark dataset D5 was created, which consists of 82 cis-Golgi proteins and 1065 trans-Golgi proteins for training. D5 was included in the Supplementary Materials. For real application, we download Human sub-Golgi proteome sequences from UniProtKB_2020.05 by using key words listed in Supplementary Table S1 footnote b.

2.2 Feature representation

Here we used protein sequence unified representation (UniRep) (Alley et al., 2019) to convert protein sequences into feature vectors. UniRep was proposed in 2019 and uses a totally different feature representation method than the previously and extensively used protein sequence feature extraction methods (Ahmad et al., 2017, 2019; Lv et al., 2019b; Rahman et al., 2018; Yang et al., 2016b; Zhou et al., 2019). Details of UniRep are found in reference (Alley et al., 2019). Briefly, UniRep used a slightly modified version of the original multiplicative long short-term memory architecture (mLSTM) (Krause et al., 2016) in Figure 1 as deep representation learner to self-supervised learning for prediction of the next following amino acid in a sequence. UniRep50 downloaded from UniProt was used to train UniRep to perform next amino-acid prediction in the protein sequences (Bateman et al., 2019). The protein feature vectors derived from UniRep have been used as the input for secondary structure, stability, diverse functions and semantic similarity of protein prediction or clustering, resulting in more efficiency for protein engineering tasks (Alley et al., 2019; Qi et al., 2020). In this work, each of sub-Golgi-localized protein sequences were firstly converted into an integer sequence according to following function:
\[ f(s_i) = i, \]
\[ i = 1, 2, \ldots, 21, \text{if } s_i \in 20 \text{ canonical amino acid} \]
\[ i = 22, \text{if } s_i \in \{O\} \]
\[ i = 23, \text{if } s_i \in \{X, B, Z, J\} \]

where \( s_i \) is the \( i \)-th amino acid of the sequence, and \( s_i \) is one of the canonical and non-canonical amino-acid symbols \([X, B, Z, J]\). The integer sequence \( f(s_i) \), \( j = 1, 2, \ldots, L \) (length of protein sequence) was embedded into 1900-long feature vectors via the UniRep method for future supervised prediction. The 1900 dimension features are calculated from the average output hidden states \( h_t \) in the UniRep model. The UniRep calculation details please see Alley et al. (2019). For comparison, some preprint works of the state-of-the-art of protein sequence deep representation embed features including TAPE from Rao et al. (2019) and BiLSTM embedding from Bepler et al. (2019) were used. See the Supplementary Text S1 for the details.

### 2.3 Feature selection

Compared to 304 training samples in D3, there are too many redundant features for a 1900-dimension UniRep feature vector of each
protein sequence, which would result in overfitting of the machine learning model. In this study, three feature selection techniques were proposed to filter valid features for sub-Golgi classification. The first was ANOVA, which sorted features by measuring the ratio of their variance between and within groups (Blanca et al., 2017; Tang et al., 2016). ANOVA has been widely used in bioinformatics, medical research and other fields (Jung et al., 2019; Tavakkolkhah et al., 2018). The second was the LGBM algorithm (Ke et al., 2017), which selected the best feature space based on feature importance values calculated by the LGBM model. Recently, LGBM feature selection methods were well applied to RNA pseudouridine site and DNA N-4-Methylcytosine sites prediction (Lv et al., 2020a; Lv et al., 2020b). LGBM is available at https://lightgbm.readthedocs.io. The third was max relevance max distance(MRMD) method (http://lab.malab.cn/soft/MRMD3.0/index.html) (Zou et al., 2016), which is an integrated tool based PageRank strategy to use different popular feature ranking algorithms to determine the properly reduced feature space.

2.4 Imbalanced data processing
The training benchmark dataset D3 and D5 are class imbalanced datasets, in which the number of cis-Golgi proteins is much less than that of trans-Golgi proteins. Reported sub-Golgi classifiers have demonstrated such class imbalance to significantly impact real application performance (Lv et al., 2019b). That is, training results are more likely to identify the majority while ignoring the minority of classes. To overcome this and depending on sufficient quantity, under-sampling is used when sufficient sample is available and over-sampling is used when insufficient sample is available (Fernandez et al., 2018). In the case of sub-Golgi classification, over-sampling is often applied using SMOTE (Barua et al., 2014), which has been integrated as a module in the imbalanced-learn toolkit (https://github.com/scikit-learn-contrib/imbalanced-learn) (Lemaitre et al., 2017).

2.5 Classifiers
In order to determine the most suitable machine learning algorithm, we tested ten popular machine learning algorithms. They were Logistic Regression (LR), KNN, Decision Tree (DT), Gaussian Naive Bayes (NB), Bagging, Random Forest (RF) (Shi et al., 2019; Wang et al., 2019a, 2020b), Ada Boosting (AB), Light Gradient Boosting Machine (LGBM), Support Vector Machine (SVM) (Huo et al., 2020; Wang et al., 2019b) and Linear Discriminant Analysis (LDA). They are out-of-the-box tools in the scikit-learn toolkit (https://github.com/scikit-learn) (Pedregosa et al., 2011). The default hyper-parameters were used for first-round classifier filtering and the top 2 classifiers were selected for hyper-parameter optimization to determine the optimal classifier. Please see the detail in the Supplementary Text S2.

2.6 Evaluation metrics and methods
As for most binary classification machine learning methods, four standard metrics including accuracy (ACC), sensitivity (Sn), specificity (Sp) and Matthew correlation coefficient (MCC) were adopted to evaluate the performance of the trained models (Ding et al., 2017, 2019; Hong et al., 2020; Li et al., 2020; Yang et al., 2018, 2019a; Zeng et al., 2018, 2019). They were calculated as equation (2) to (5). For TP, TN, FP and FN when short of the predicted sample number of true positive, true negative, false positive and false negative. In this study, we denoted proteins at the cis-Golgi location as positive samples and proteins at the trans-Golgi location as negative samples.

\[
ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)
\]

\[
Sn = \frac{TP}{TP + FN} \quad (3)
\]

\[
Sp = \frac{TN}{TN + FP} \quad (4)
\]

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (5)
\]

The area value (auROC) under the receiver operating characteristic (ROC) curve was also applied for model evaluation (Dao et al., 2020a,b; Fung et al., 2019; Zhang et al., 2008). ROC was constructed by plotting the true positive rate with respect to the false positive rate, with auROC ranging from 0 to 1. An auROC of 1 indicated perfect prediction while an auROC of 0.5 indicated random predictions for both positive and negative samples.

In addition to the evaluation metrics, 10-fold and leave-one-out (LOO) cross validation (Deng et al., 2020; Zhang et al., 2019a,b) and independent testing protocols were utilized (Jiao et al., 2016a). Please see the detail in the Supplementary Text S3.

3 Results
3.1 Initial performances of different classifiers
First, we used ten widely used machine learning algorithms to classify cis-Golgi and trans-Golgi protein sequences represented by the 1900-dimension deep learning features without data balancing or feature selection. The 10-fold cross-validation accuracy box plotting and ROC curves of these ten classifiers are shown in Figure 2A and B. The support vector machine classifier with the best average accuracy was 77.3% with an auROC value of 0.765, and an MCC value of 0.379 (see Supplementary Table S2). The SVM classifier was again top ranked for its high accuracy and auROC value.

The training dataset D3 contains 304 sequence with 87 cis-Golgi located protein sequence and 217 trains-Golgi located protein sequences. Obviously, the 1900 dimensional features are far more than the training sample numbers that would cause feature dimension redundancy and potential overfitting, so for the third step, we used LGBM feature selection technology to sort the 1900 deep representation learning features by their importance value with selection of the top 250 features for model fitting. The results are shown in Figure 2E, F and Supplementary Table S2. After feature dimension reduction, performance of some machine learning algorithms declined. For example, the accuracy of Gaussian Naive Bayes (NB), LR and AB decreased by the absolute value of 14%, 0.7% and 0.85%, respectively. The accuracy of the other seven classifiers rose by the absolute values ranging from 0.4% to 7.5%, among which SVM ranked top at ACC = 90.76%, LGBM ranked second best at ACC = 90.57%. Since the performance SVM and LGBM were quite close, both were chosen for subsequent optimization and comparison.

3.2 Effect of feature selection technologies on SVM and LGBM classifiers
To determine the optimal feature space for SVM and LGBM classifiers, we used a two-step feature optimizing strategy. In the first step, we used three feature selection technologies, (ANOVA, MRMD and LGBM) to calculate the feature importance values like F-score for ANOVA, page-ranked values for MRMD and Gini-based feature importance values for LGBM to yield three descending order lists for deep representation learning features. In the second step, for each feature list obtained by different methods, we selected the top 200 features and used a feature-by-feature incremental strategy to determine the optimal feature vector space for SVM and LGBM classifiers. The 10-fold cross-validation accuracy results are shown in Figure 3A. For example, the black curve with the ANOVA_SVM legend is the accuracy of the SVM model varying by the increasing number of features selected by the ANOVA method.
Fig. 2. Ten-fold cross-validation accuracy metrics of Boxplots and ROC curves for ten classifiers (LR: Logistic Regression, KNN: K-nearest Neighbors, DT: Decision Tree, NB: Gaussian Naive Bayes, Bagging: Bagging, RF: Random Forest, AB: Ada Boosting, LGBM: Light Gradient Boosting Machine, SVM: Supporting Vector Machine, LDA: Linear Discriminant Analysis) using different feature processing technologies. A and B utilized UniRep feature vectors with 1900 dimensions; C and D used SMOTE to balance the UniRep feature vectors with 1900 dimensions; for E and F, based on the previous steps, 250 features were selected by using the LGBM feature selection method. Green Triangles and orange lines in A, C and E are the average accuracy values and the median accuracy values for the 10-fold cross-validation. In either case, SVM classifier had the highest average accuracy (77.32%, 90.31% and 90.76%, respectively) and the highest average auROC value (0.765, 0.940 and 0.958, respectively)
Fig. 3. (A) Based on benchmark dataset D3, the average 10-fold cross-validation accuracy varied with the feature numbers for LGBM and SVM classifiers based on ANOVA, MRMD, and LGBM feature selection technology. The best SVM had an accuracy of 92.16% with 158 features. The best LGBM classifier had an accuracy of 93.08% with 64 features. Both were based on LGBM feature selection technology. (B) Ten-fold cross-validation and LOO metrics for comparison of the best SVM (based on benchmark dataset D3 and D5) and LGBM classifier (based on benchmark dataset D3). (C) Independent test metrics on benchmark testing dataset D4 for the best SVM and LGBM classifier obtained by LOO using benchmark dataset D3 and D5.
Initially, as the number of features increased, each accuracy curve for the model increased sharply and then approached fluctuating plateaus. For the accuracy value of each curve plateau, both curves as well as the violin-box plotting insets, which are statistical ods. The maximum accuracy of the LGBM_LGBM model was 93.08% with 64 features and was more than the 92.16% for the LGBM_SVM model with 158 features. The accuracy of the LGBM_LGBM model was more accurate than those based on ANOVA and MRMD methods. The fluctuating purple and golden curves for the model increased sharply and then approached fluctuat-

### Table 1: Evaluation metrics comparisons of support vector machine classifiers based on different state-of-the-art deep representation learning features

| Feature type | Trained dataset (identity) | Feature dimensions | LOO cross-validation | Independent testing |
|--------------|----------------------------|--------------------|----------------------|---------------------|
|              |                            |                    | ACC (%) MCC Sn (%) Sp (%) auROC | ACC (%) MCC Sn (%) Sp (%) auROC |
| UniRep       | D3 (40%)                   | 158                | 92.6 0.85 94.9 90.3 0.964 | 98.4 0.95 100 98.0 0.995 |
|              | D5 (25%)                   | 107                | 99.2 0.98 100 98.4 0.999 | 96.4 0.90 100 84.6 0.994 |
| BiLSTM-lm    | D3 (40%)                   | 77                 | 88.7 0.78 85.2 92.1 0.917 | 92.1 0.75 96.1 76.9 0.983 |
|              | D5 (25%)                   | 152                | 99.4 0.99 99.9 98.9 0.999 | 92.2 0.75 100 61.5 0.989 |
| BiLSTM-ssa   | D3 (40%)                   | 93                 | 91.7 0.83 90.7 92.6 0.946 | 90.6 0.71 94.1 76.9 0.975 |
|              | D5 (25%)                   | 48                 | 99.8 0.99 100 99.7 0.999 | 87.5 0.58 100 38.5 0.956 |
| TAPE-pooled  | D3 (40%)                   | 77                 | 90.3 0.81 89.9 90.7 0.941 | 90.6 0.70 96.0 69.2 0.966 |
|              | D5 (25%)                   | 53                 | 98.7 0.97 100 97.5 0.999 | 90.6 0.69 98.0 61.5 0.927 |
| TAPE-avg     | D3 (40%)                   | 67                 | 91.9 0.84 94.0 89.9 0.963 | 96.4 0.91 100 96.1 0.985 |
|              | D5 (25%)                   | 73                 | 99.7 0.99 100 99.3 0.999 | 89.1 0.64 100 46.1 0.989 |

### 3.3 Comparison with the state-of-the-art deep representation feature types

For comparison, four types of deep representation feature from two preprint works were used for selection of best feature type for the sub-Golgi prediction task. The leave-one-out cross-validation and testing results of SVM model trained on D3 and D5 dataset using BiLSTM-lm, BiLSTM-ssa, TAPE-pooled and TAPE-avg features respectively are listed in Table 1. For models trained on D3, the model using UniRep features obtained the best leave-one-out cross-validation accuracy scores and independent testing scores. For models trained on D5, the model using UniRep features got the value 99.2% of leave-one-out cross-validation accuracy, a little smaller than the value of models using BiLSTM-lm, BiLSTM-ssa, TAPE-avg by the relative values of 0.2%, 0.6% and 0.5%, separately. While the independent testing accuracy of model using UniRep feature is 96.4%, much greater than the value of models using BiLSTM-lm, BiLSTM-ssa, TAPE-avg by the relative values of 4.5%, 10.1% and 8.2%.

### 3.4 Comparison with the state-of-the-art classifiers

To further evaluate the performance of our classifiers, we compared isGP-DRLF with state-of-the-art classifiers in Supplement Table S3. Supplementary Table S3 summarizes the LOO cross-validation metrics of published classifiers based on different benchmark training dataset D0, D1, D2 and D3 listed in Supplement Table S1. Classifiers based on the D3 dataset were superior to other classifiers based on D0, D1 and D2. Apart from the isGP-DRLF of this study and Ding’s SVM model (Ding et al., 2013), other reported sub-Golgi classifiers had to fuse multi-type features to attain acceptable results. The isGP-DRLF was far better than Ding’s SVM model (Ding et al., 2013) as judged by LOO cross-validation (Supplementary Table S3)
To explain the difference between the models and feature representation capability, we used UMAP (uniform manifold approximation and projection) method (McInnes et al., 2018) to reduce the UniRep feature space and 2 gap dipeptide composition features space dimensions into two and the dimension reduction results were shown in in the Supplementary Figure S3. From Supplementary Figure S3A and B, it found out that it was better to use UniRep features than use 2gDPC features for identifying proteins located in cis-Golgi from proteins located in trans-Golgi. Also, as displayed in Supplementary Figure S3A to H for protein sub-Golgi localization task, UniRep features are superior to some widely used feature types listed in Supplementary Table S3. The strong sequence feature representation capability of UniRep enables us to use only one feature type to achieve good classification accuracy without using features fusion technology.

3.5 Webserver implementation

The isGP-DRLF is now available at http://isGP-DRLF.aibiochem.net and its interface is shown in Supplementary Figure S4. The webserver allows input of FASTA format protein sequences and identifies whether a protein is found in cis-Golgi or trans-Golgi. The users could paste FASTA format Golgi protein sequence in the left blank box and click the submit button to calculate. After a while, the prediction results will be showed in the right table. Before starting a new task, the users have to clear the input box first to reactivate the submit button. Due to limited computing resources, do not input more than 5 sequences at a time. For larger dataset computing, the users could download the python script and the trained model from https://github.com/zibsniv/isGP-DRLF.

4 Conclusion

A novel state-of-the-art sub-Golgi protein localization classifier, isGP-DRLF, was developed by use of protein sequence deep representation learning feature vectors. Combined with over-sampling, imbalanced data processing and LGBM feature selection, isGP-DRLF achieved the best independent testing evaluation accuracy values (ACC = 96.4%) for sub-Golgi protein prediction, which increases by 1.15% relative to the corresponding best value (ACC = 95.3%) for the previously reported models isGPT (Rahman et al., 2018) listed in Table 2. For independent testing error rate, the absolute error rate value of isGP-DRLF is 3.6%, which reduced by 23.4% relative to the absolute error rate value 4.7% of the previous best-reported model isGPT (Rahman et al., 2018). The isGP-DRLF employs just one type of sequence representation feature with performance superior to other state-of-the-art sub-Golgi classifiers fusing multiple types of feature representations. This study shows that protein sequence deep representation learning is a type of very discriminating feature vector that distinguishes multiple sequences, which will be useful in the future for a more accurate prediction of protein multi-subcellular localization (Armenteros et al., 2017) and for recognition of protein chemical modification sites, signal peptide without a requirement for multi-type feature fusion (Armenteros et al., 2019; Yang et al., 2016a). The current model only applies to sub-Golgi prediction. In the future, we will apply deep presentation

---

**Table 2. Evaluation metrics comparisons of the state-of-the-art classifiers**

| Classifier              | Trained dataset | Feature type numbers | Features dimensions | LOO Cross-validation | Independent testing |
|-------------------------|-----------------|----------------------|---------------------|----------------------|---------------------|
| SVM (this study)        | D3              | 1                    | 158                 | ACC: 92.6 MCC: 0.85  | ACC: 98.4 MCC: 0.95 |
| SVM (this study)        | D5              | 1                    | 107                 | ACC: 99.2 MCC: 1.00  | ACC: 96.4 MCC: 0.90 |
| KNN (Ahmad et al., 2017)| D3              | 3                    | 83                  | ACC: 94.9 MCC: 0.90  | ACC: 94.8 MCC: 0.86 |
| KNN (Ahmad et al., 2019)| D3              | 3                    | 180                 | ACC: 98.2 MCC: 0.96  | ACC: 94.0 MCC: 0.84 |
| RF (Yang et al., 2016b) | D3              | 4                    | 55                  | ACC: 88.5 MCC: 0.68  | ACC: 93.8 MCC: 0.82 |
| SVM (Rahman et al., 2018)| D3              | 6                    | 2800                | ACC: 95.9 MCC: 0.92  | ACC: 95.3 MCC: 0.85 |

Fig. 4. Human sub-Golgi proteome sequence distribution and the results of isGP-DRLF and subGolgi2 tested on human sub-Golgi proteome dataset and independent testing scores (Supplementary Table S4). The six models based on the D3 training are shown in Supplement Tables S3 and S4 and in the main text Table 2. KNN sub-Golgi localization protein classifier developed by Ahmad et al. (2019) had Split Amino Acid Composition (SAAC), 3gap Dipeptide Composition (3gDPC) and Position Specific Scoring Matrix (PSSM) fusion features and realized a best LOO accuracy of 98.2% with independent test accuracy of 94.0%. Our isGP-DRLF based on D5 had 107 UniRep features and achieved 99.2% LOO accuracy and independent test accuracy of 96.4%. Evidently, given the independent testing results of the models based on D3 and D5 in this study, isGP-DRLF is better at predicting unknown sub-Golgi protein sequences.

Considering the independent testing dataset D4 consisting of 64 sequence and to test the model practical application capacity, we also applied isGP-DRLF to the human sub-Golgi proteome dataset with 423 sequence. The sequence location distribution is drawn in Figure 4A. It attained 91.8% accuracy for reviewed human sub-Golgi proteome while it predicted that about 28% unreviewed human sub-Golgi proteome would be located in cis-Golgi and 72% would be located in trans-Golgi (see Fig. 4B and D). While among state-of-the-art predictors listed in Table 2, now only Lin’s subGolgi2 webserver is available (http://lin-group.cn/server/subGolgi2) (Ding et al., 2013). The subGolgi2 is a predictor using 2 gap dipeptide composition (2gDPC) features and we also tested it with human sub-Golgi proteome dataset and results are shown in Figure 4C and D. For reviewed sequences, subGolgi2 got accuracy of 77.3%, by the relative value of 16% lower than that of isGP-DRLF.

---

Sub-Golgi proteins localization

To explain the difference between the models and feature representation capability, we used UMAP (uniform manifold approximation and projection) method (McInnes et al., 2018) to reduce the UniRep feature space and 2 gap dipeptide composition features space dimensions into two and the dimension reduction results were shown in in the Supplementary Figure S3. From Supplementary Figure S3A and B, it found out that it was better to use UniRep features than use 2gDPC features for identifying proteins located in cis-Golgi from proteins located in trans-Golgi. Also, as displayed in Supplementary Figure S3A to H for protein sub-Golgi localization task, UniRep features are superior to some widely used feature types listed in Supplementary Table S3. The strong sequence feature representation capability of UniRep enables us to use only one feature type to achieve good classification accuracy without using features fusion technology.

3.5 Webserver implementation

The isGP-DRLF is now available at http://isGP-DRLF.aibiochem.net and its interface is shown in Supplementary Figure S4. The webserver allows input of FASTA format protein sequences and identifies whether a protein is found in cis-Golgi or trans-Golgi. The users could paste FASTA format Golgi protein sequence in the left blank box and click the submit button to calculate. After a while, the prediction results will be showed in the right table. Before starting a new task, the users have to clear the input box first to reactivate the submit button. Due to limited computing resources, do not input more than 5 sequences at a time. For larger dataset computing, the users could download the python script and the trained model from https://github.com/zibsniv/isGP-DRLF.

4 Conclusion

A novel state-of-the-art sub-Golgi protein localization classifier, isGP-DRLF, was developed by use of protein sequence deep representation learning feature vectors. Combined with over-sampling, imbalanced data processing and LGBM feature selection, isGP-DRLF achieved the best independent testing evaluation accuracy values (ACC = 96.4%) for sub-Golgi protein prediction, which increases by 1.15% relative to the corresponding best value (ACC = 95.3%) for the previously reported models isGPT (Rahman et al., 2018) listed in Table 2. For independent testing error rate, the absolute error rate value of isGP-DRLF is 3.6%, which reduced by 23.4% relative to the absolute error rate value 4.7% of the previous best-reported model isGPT (Rahman et al., 2018). The isGP-DRLF employs just one type of sequence representation feature with performance superior to other state-of-the-art sub-Golgi classifiers fusing multiple types of feature representations. This study shows that protein sequence deep representation learning is a type of very discriminating feature vector that distinguishes multiple sequences, which will be useful in the future for a more accurate prediction of protein multi-subcellular localization (Armenteros et al., 2017) and for recognition of protein chemical modification sites, signal peptide without a requirement for multi-type feature fusion (Armenteros et al., 2019; Yang et al., 2016a). The current model only applies to sub-Golgi prediction. In the future, we will apply deep presentation
learning features for eukaryotic proteins multiple subcellular and suborganellar localization prediction, functioning like DeepLoc (Armenteros et al., 2017) or MULoCDeep (Jiang et al., 2020a).

Acknowledgements

The authors are really appreciated to the three anonymous reviewers. Their constructive comments are very helpful for strengthening the presentation of this paper.

Funding

The work was funded by the National Natural Science Foundation of China [62001080, 91935302, 61922020, 61822108 and 61771331], and by the China Postdoctoral Science Foundation [2020M673184].

Conflict of Interest: none declared.

References

Ahmad,J. et al. (2019) MFSC: multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. J. Theor. Biol., 463, 99–109.

Ahmad,J. et al. (2017) Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods. Artif. Intell. Med., 78, 14–22.

Alley,E.C. et al. (2019) Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods, 16, 1315–1322.

Armenteros,J.J.A. et al. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics, 33, 4049–4049.

Armenteros,J.J.A. et al. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol., 37, 420.

Barua,S. et al. (2014) MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowl. Data Eng., 26, 405–425.

Bateman,A. et al. (2019) UniProt: a worldwide hub of protein knowledge. Nucl. Acids Res., 47, D506–D515.

Bepler,T. et al. (2019) Learning protein sequence embeddings using information from structure. arXiv:1902.08664.

Berry,K.P. et al. (2017) Spine dynamics: are they all the same? Neuron, 96, 43–55.

Blanca,M.J. et al. (2017) Non-normal data: is ANOVA still a valid option? Psicothema, 29, 552–557.

Dao,F.Y. et al. (2020a) Computational identification of N6-methyladenosine sites in multiple tissues of mammals. Comput. Struct. Biotechnol. J., 18, 1084–1091.

Dao,F.Y. et al. (2020b) A computational platform to identify origins of replication sites in eukaryotes. Brief Bioinf., 10.1093/bib/bbaa1017.

De Tito,S. et al. (2020) The Golgi as an Assembly Line to the Autophagosome. Trends Biochem. Sci., 45, 484–496.

Deng,Y. et al. (2020) A multimodal deep leaning framework for predicting drug-drug interaction events. Bioinformatics, 36, 4316–4322.

Ding,H. et al. (2013) Prediction of Golgi-resident protein types by using feature selection technique. Chemom. Intell. Lab. Syst., 124, 9–13.

Ding,H. et al. (2011) Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. Protein Peptide Lett., 18, 58–63.

Ding,H. et al. (2016) PHYPred: a tool for identifying bacteriofage enzymes and hydrolases. Virologica Sinica, 31, 330–352.

Ding,Y. et al. (2017) Identification of drug-target interactions via multiple information integration. Inf. Sci., 418–419, 546–560.

Ding,Y. et al. (2019) Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing, 325, 211–224.

Eraslan,G. et al. (2019) Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet., 20, 389–403.

Feng,C.Q. et al. (2019) iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics, 35, 1469–1477.

Fernandez,A. et al. (2018) SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res., 61, 863–905.

Fujita,Y. et al. (2006) Fragmentation of Golgi apparatus of nigral neurons with alpha-synuclein-positive inclusions in patients with Parkinson’s disease. Acta Neuropathol., 112, 261–265.

Gonatas,N.K. et al. (1998) The involvement of the Golgi apparatus in the pathogenesis of amyotrophic lateral sclerosis, Alzheimer’s disease, and ricin intoxication. Histochem. Cell Biol., 109, 591–600.

Holthuis,J.C.M. et al. (2014) Lipid landscapes and pipelines in membrane homeostasis. Nature, 510, 48–57.

Hong,Z. et al. (2020) Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics, 36, 1037–1043.

Huo,Y. et al. (2020) SGL-SVM: a novel method for tumor classification via support vector machine with sparse group Lasso. J. Theor. Biol., 486, 1100098.

Jiang,Y. et al. (2019a) A dynamic programing approach to integrate gene expression data and network information for pathway model generation. Bioinformatics, 36, 169–176.

Jiang,Y. et al. (2019b) DeepDom: predicting protein domain boundary from sequence alone using stacked bidirectional LSTM. In: Ahman,R.B. et al. (eds.) Pacific Symposium on Biocomputing 2019 pp. 66–75.

Jiang,Y. et al. (2020a) MULoCDeep: a deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. doi: 10.21203/rs.21203.v4.7047.121001.

Jiang,Y. et al. (2020b) IMPRes-Pro: a high dimensional multiotics integration method for in silico hypothesis generation. Methods, 173, 16–23.

Jiao,Y. et al. (2016a) Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant. Biol., 4, 320–330.

Jiao,Y. et al. (2016b) Predicting Golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties. J. Theor. Biol., 391, 35–42.

Jiao,Y. et al. (2016c) Prediction of Golgi-resident protein types using general form of Chou’s pseudo-amino acid compositions: approaches with minimal redundancy maximal relevance feature selection. J. Theor. Biol., 402, 38–44.

Jung,Y. et al. (2019) Transformed low-rank ANOVA models for high-dimensional variable selection. Stat. Methods Med. Res., 28, 1230–1246.

Ke,G.L. et al. (2017) LightGBM: a highly efficient gradient boosting decision tree. In: Guyon,I. et al. (eds.) Advances in Neural Information Processing Systems 30, Neural Information Processing Systems (NIPS), La Jolla.

Krause,J. et al. (2016) Multiplicative LSTM for sequence modelling. arXiv e-Prints, arXiv:1609.07959.

Lemaitre,G. et al. (2017) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res., 18, 5.

Li,J. et al. (2020) DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. IEEE J. Biomed. Health Inf., 24, 3012–3019.

Ligon,C. et al. (2020) A selective role for a component of the autophagy pathway in coupling the Golgi apparatus to dendrite polarity in pyramidal neurons. Neurosci. Lett., 730, 7.

Lv,Z. et al. (2019a) Protein function prediction: from traditional classifier to deep learning. Proteomics, 19, 1900119.

Lv,Z. et al. (2019b) A random forest sub-golgi protein classifier optimized via deep peptide and amino acid composition features. Front. Bioeng. Biotechnol., 7, 215.

Lv,Z. et al. (2020a) Escherichia coli DNA N4-methycytosine site prediction accuracy improved by light gradient boosting machine feature selection technology. IEEE Access, 8, 14851–14859.

Lv,Z. et al. (2020b) RF-PseU: a random forest predictor for DNA pseudouridine sites. Front. Bioeng. Biotechnol., 8, 134.

Mcinnes,L. et al. (2018) UMAP: uniform manifold approximation and projection for dimension reduction. J. Open Source Softw., 3, 861.

Min,S. et al. (2017) Deep learning in bioinformatics. Brief. Bioinf., 18, 851–869.

Nambari,A. et al. (2020) Transforming the language of life: transformer neural networks for protein prediction tasks. BioRxiv, 2020.2006.2015.153643.

Pedregosa,F. et al. (2011) Scikit-learn: machine learning in Python. J. Mach. Learn. Res., 12, 2825–2830.

Qi,Y. et al. (2020) DenseCPD: improving the accuracy of neural-network-based computational protein sequence design with DenseNet. J. Theor. Biol., 510, 510–525.

Rahman,M.S. et al. (2018) iGSP: an optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection. Artif. Intell. Med., 84, 90–100.
Sub-Golgi proteins localization

---

Rao, R. et al. (2019) Evaluating Protein Transfer Learning with TAPE. arXiv: 1906.08230.
Ravichandran, Y. et al. (2020) The Golgi apparatus and cell polarity: roles of the cytoskeleton, the Golgi matrix, and Golgi membranes. *Carr. Cell Biol.*, 62, 104–113.
Shen, Y. et al. (2019a) Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinf.*, 21, 1628–1640.
Shen, Y. et al. (2019b) Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou’s general PseAAC. *J. Theor. Biol.*, 462, 230–239.
Shi, H. et al. (2019) Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, 111, 1839–1852.
Stoeber, M. et al. (2018) A genetically encoded biosensor reveals location bias of opioid drug action. *Neuron*, 98, 963–976.e5.
Tang, H. et al. (2016) Identification of immunoglobulins using Chou’s pseudo amino acid composition with feature selection technique. *Mol. BioSyst.*, 12, 1269–1275.
Tang, H. et al. (2018) HBPred: a tool to identify growth hormone–binding proteins. *Int. J. Biol. Sci.*, 14, 957–964.
Tao, Y. et al. (2020) Golgi apparatus: an emerging platform for innate immunity. *Trends Cell Biol.*, 30, 467–477.
Tavakkolkhah, P. et al. (2018) Detection of network motifs using three-way ANOVA. *PLoS One*, 13, e021382.
Wang, S. et al. (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, 13, 34.
Wang, X. et al. (2019a) Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics*, 35, 2395–2402.
Wang, Y. et al. (2019b) Pancreatic cancer biomarker detection by two support vector strategies for recursive feature elimination. *Biomarkers Med.*, 13, 105–121.
Wang, H. et al. (2020a) Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion. *Neurocomputing*, 383, 257–269.
Wang, Z. et al. (2020b) Identification of highest-affinity binding sites of yeast transcription factor families. *J. Chem. Inf. Model.*, 60, 1876–1883.
Xu, J. B. (2019) Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. USA*, 116, 16856–16865.
Xu, J. B. et al. (2017) Folding Large Proteins by Ultra-Deep Learning. *Assoc Computing Machinery*, New York.
Xu, J. B. et al. (2019) Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins*, 87, 1069–1081.
Yang, A. et al. (2016a) A chemical biology route to site-specific authentic protein modifications. *Science*, 354, 623–626.
Yang, R. et al. (2016b) A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data. *Int. J. Mol. Sci.*, 17, 218–217.
Yang, H. et al. (2018) iRNA-2OM: a sequence-based predictor for identifying 2’-O-methylation sites in *Homo sapiens*. *J. Comput. Biol.*, *J. Comput. Mol. Cell Biol.*, 25, 1266–1277.
Yang, H. et al. (2019a) An unsupervised conditional graphical model for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief. Bioinf.*, 10.1093/bib/bbz1123.
Yang, W. et al. (2019b) A brief survey of machine learning methods in protein sub-golgi localization. *Carr. Biol.*, 14, 234–240.
Zeng, X. et al. (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35, 5191–5198.
Zeng, X. et al. (2018) Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics*, 34, 2425–2432.
Zhang, L. et al. (2016) LSDF: latent sparse domain transfer learning for visual adaptation. *IEEE Trans. Image Process.*, 25, 1177–1191.
Zhang, W. et al. (2019a) SFLLN: a sparse feature learning ensemble method with linear neighborhood regularlization for predicting drug–drug interactions. *Inf. Sci.*, 497, 189–201.
Zhang, W. et al. (2019b) A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, doi:10.1109/TCBB.2019.2931546.
Zhang, W. et al. (2008) A Bayesian approach to the prediction of HIC-II binding affinity. *Comput. Methods Programs Biomed.*, 92, 1–7.
Zhao, W. et al. (2019) Predicting protein sub-Golgi locations by combining functional domain enrichment scores with pseudo-amino acid composition. *J. Theor. Biol.*, 473, 38–43.
Zhou, H. Y. et al. (2019) Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion. *IEEE Access*, 7, 144154–144164.
Zhou, M. et al. (2020) Progress in neural NLP: modeling, learning, and reasoning. *Engineerings*, 6, 275–290.
Zou, Q. et al. (2016) A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing*, 173, 346–354.