Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

isGPT: An optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection

M. Saifur Rahman[a,1], Md. Khaledur Rahman[b,1], M. Kaykobad[a], M. Sohel Rahman[a,*]

[a] Department of CSE, BUET, ECE Building, West Palasi, Dhaka 1205, Bangladesh
[b] Indiana University, Bloomington, USA

ARTICLE INFO

Article history:
Received 5 October 2017
Received in revised form
13 November 2017
Accepted 17 November 2017

Keywords:
Sub-Golgi Apparatus
Classification
Regression
Support vector machine
Random Forest

ABSTRACT

The Golgi Apparatus (GA) is a key organelle for protein synthesis within the eukaryotic cell. The main task of GA is to modify and sort proteins for transport throughout the cell. Proteins permeate through the GA on the ER (Endoplasmic Reticulum) facing side (*cis* side) and depart on the other side (*trans* side). Based on this phenomenon, we get two types of GA proteins, namely, *cis*-Golgi protein and *trans*-Golgi protein. Any dysfunction of GA proteins can result in congenital glycosylation disorders and some other forms of difficulties that may lead to neurodegenerative and inherited diseases like diabetes, cancer and cystic fibrosis. So, the exact classification of GA proteins may contribute to drug development which will further help in medication.

In this paper, we focus on building a new computational model that not only introduces easy ways to extract features from protein sequences but also optimizes classification of *trans*-Golgi and *cis*-Golgi proteins. After feature extraction, we have employed Random Forest (RF) model to rank the features based on the importance score obtained from it. After selecting the top ranked features, we have applied Support Vector Machine (SVM) to classify the sub-Golgi proteins. We have trained regression model as well as classification model and found the former to be superior. The model shows improved performance over all previous methods. As the benchmark dataset is significantly imbalanced, we have applied Synthetic Minority Over-sampling Technique (SMOTE) to the dataset to make it balanced and have conducted experiments on both versions. Our method, namely, *identification of sub-Golgi Protein Types (isGPT)*, achieves accuracy values of 95.4%, 95.9% and 95.3% for 10-fold cross-validation test, jack-knife test and independent test respectively. According to different performance metrics, isGPT performs better than state-of-the-art techniques. The source code of isGPT, along with relevant dataset and detailed experimental results, can be found at https://github.com/srautonu/isGPT.

1. Introduction

An eukaryotic cell is defined by a membrane-bound nucleus. All eukaryotic cells have a nucleus, a plasma membrane, ribosomes and cytoplasm [1]. Most of the eukaryotic cells have other small membrane-bound structures in cytoplasm called organelles and Golgi Apparatus (GA) is one of them. It is a key organelle in protein synthesis along with some other elements of the cell [2]. It consists of disk like membranes called cisternae which are stacked together [3]. GA has three elements, namely, *cis*-Golgi, medial and *trans*-Golgi. *cis*-Golgi is responsible for receiving proteins, while *trans*-Golgi releases the synthesized proteins. The function of medial is to synthesize the received proteins from *cis*-Golgi (see Fig. 1, image source: [4]). Endoplasmic Reticulum (ER) builds proteins and sends out to the cell through GA [4]. A side of GA facing ER (*cis*-side) captures those proteins (also called cargo proteins) for synthesis and send those out via the other side of GA facing the plasma membrane (*trans*-side). In the medial region, the cargo proteins get modified by the Golgi enzymes through addition or removal of sugars. Modifications may also occur through the addition of sulphate groups or phosphate groups [4].

Any functional deviation of GA may result in adaptable disorders during the synthesis process in medial which may further contribute to inheritable and neurodegenerative diseases such as diabetes [5], cancer [6], Parkinson's disease [7] and Alzheimer's disease [8]. It is necessary to identify any rambling and damage

[*] Corresponding author.
E-mail addresses: mrahman@cse.buet.ac.bd (M.S. Rahman), morahma@iu.edu (Md.K. Rahman), kaykobad@cse.buet.ac.bd (M. Kaykobad), msrahman@cse.buet.ac.bd (M.S. Rahman).
[1] Equal contributors.

https://doi.org/10.1016/j.artmed.2017.11.003
0933-3657/© 2017 Elsevier B.V. All rights reserved.

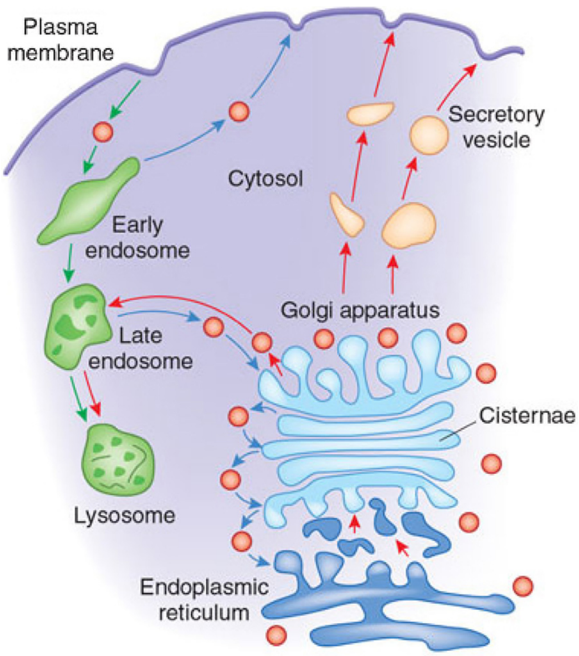

Fig. 1. The Golgi Apparatus and its synthesis process.
Image source: [4].

in a timely manner to better understand the problem of GA dysfunction. The current methods of treating patients having these diseases include neuroprotective therapies and anti-inflammation which are not able to provide a permanent solution [9]. Exact identification of sub-Golgi proteins can provide new insight for scientists to recognize the dysfunctions subscribed by Golgi proteins [10]. Thus, sub-Golgi (*cis*-Golgi vs. *trans*-Golgi) protein classification is very important for more effective drug-development.

Significant amount of research have been conducted during the last decade to build prediction tools for protein sub-cellular localization using machine learning methods [11–16]. However, very few tools have been developed for sub-Golgi protein classification. Nevertheless, researchers nowadays are focusing on this topic and trying to build efficient classification models. van Dijk et al. did the pioneering work to predict sub-Golgi localization of type II membrane proteins [17]. They used amino acid grouping in conjunction with string-based triads as well as 3D-structure based triads for protein representation. Here, Support Vector Machine (SVM) [18] with linear kernel was used as the classifier.

Ding et al. used increment of Shannon entropy (IH) on Amino Acid Compositions (AAC) and g-gap dipeptide compositions for protein representations. They then applied Modified Mahalanobis Discriminant (MD) algorithm to predict the Golgi-resident proteins [19]. They achieved an accuracy of 74.7% using jackknife cross-validation test. Ding et al. further continued their previous work and proposed a g-gap dipeptide based feature extraction technique [20]. They used analysis of variance (ANOVA) test to select relevant features and applied SVM as the learner. This time, they obtained an accuracy of 85.4% using jackknife cross-validation.

Jiao et al. [21] presented a model which computes Positional Specific Physico-Chemical Properties (PSPCP) of a protein sequence. The PSPCP essentially integrates the Position Specific Scoring Matrix (PSSM) with the different physicochemical property values. ANOVA was applied for feature selection, while SVM with RBF kernel was used as the learner, to achieve an accuracy of 86.9%. In a subsequent study [22], they applied minimum Redundancy Maximum Relevance (mRMR) feature selection technique, instead of ANOVA, on the same feature space. This improved the accuracy to 91%.

Both Ding et al. and Jiao et al. used a small benchmark dataset, where there are only 150 GA proteins. In addition, the dataset is highly imbalanced - the number of *trans*-Golgi proteins is significantly lower than the number of *cis*-Golgi proteins. Yang et al. have recently created an updated benchmark dataset where there are 304 sub-Golgi proteins for training and 64 sub-Golgi proteins for testing the classification model [23]. They applied Synthetic Minority Over-sampling Technique (SMOTE) [24] to balance the dataset. They conducted experiments on both the imbalanced and balanced datasets and demonstrated improved accuracy with the balanced version. For feature selection, they used Random Forest (RF) [25] based recursive feature elimination method. They then applied RF as the learning method as well. Their model shows an accuracy of 88.5%, 93.8% and 90.1% for jackknife cross-validation, independent testing and 10-fold cross-validation, respectively.

Very recently, Ahmad et al. have also conducted similar kind of experiments though their feature construction, feature selection and learning algorithms are different [26]. They have applied fisher feature selection method to select relevant features and k-nearest neighbor (KNN) algorithm as the learner. The model proposed by Ahmad et al. reports an accuracy of 94.9%, 94.8% and 94.9% on the balanced benchmark dataset for jackknife cross-validation, independent testing and 10-fold cross-validation, respectively.

Exploring previous studies, we note that there is still room for improvement because even a small improvement in accuracy is highly demanding in bioinformatics tools. Improved accuracy can also contribute to better drug-development which is maintained by sensible computer-aided design [27,28].

There are three important tasks in the pathway of protein function predictions [29]. These include processing of datasets, construction of features from protein sequences and application of a suitable classification algorithm. In this paper, we first construct a large set of features based on three feature construction techniques and then apply Random Forest (RF) algorithm on the constructed feature set. We select relevant features based on the importance score provided by the RF model. Then, we apply SVM on the selected features for both classification and regression analyses. Our tool, named *isGPT*, is evaluated based on several well-established performance metrics and demonstrates superiority over existing methods.

Our overall contributions are summarized as follows:

- We present an easy and flexible method that produces several position specific as well as position independent features from protein sequences. Then feature selection is performed based on the importance score provided by the RF model.
- We model the problem of sub-Golgi protein localization both as a classification problem and a regression problem. Using SVM, we train classification models as well as regression models on the benchmark (imbalanced) dataset as well as on the dataset, balanced with a celebrated balancing technique called SMOTE. We make a comparative analysis of the different models.
- Finally, through extensive experiments, we compare isGPT with the methods of [23,26] which are currently two state-of-the-art techniques. Our method shows superior results according to different performance metrics.

2. Methods

2.1. Dataset

We have collected the training and testing benchmark datasets from Yang et al. [23], which have also been used by Ahmad et al.

Please cite this article in press as: Rahman MS, et al. isGPT: An optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection. Artif Intell Med (2017), https://doi.org/10.1016/j.artmed.2017.11.003