Journal of Theoretical Biology 402 (2016) 38–44

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Prediction of Golgi-resident protein types using general form of Chou's pseudo-amino acid compositions: Approaches with minimal redundancy maximal relevance feature selection

Ya-Sen Jiao, Pu-Feng Du *

School of Computer Science and Technology, Tianjin University, Tianjin 300354, China

HIGHLIGHTS

- A novel general form pseudo-amino acid composition.
- An effective feature selection method to find minimal feature set to predict Golgi-protein types.
- Representing the protein sequence by incorporating evolutionary information.

ARTICLE INFO

Article history:
Received 19 March 2016
Received in revised form
19 April 2016
Accepted 26 April 2016
Available online 4 May 2016

Keywords:
Golgi-resident proteins
mRMR
PseAAC
PSPCP
Feature selection
SVM

ABSTRACT

Recently, several efforts have been made in predicting Golgi-resident proteins. However, it is still a challenging task to identify the type of a Golgi-resident protein. Precise prediction of the type of a Golgi-resident protein plays a key role in understanding its molecular functions in various biological processes. In this paper, we proposed to use a mutual information based feature selection scheme with the general form Chou's pseudo-amino acid compositions to predict the Golgi-resident protein types. The positional specific physicochemical properties were applied in the Chou's pseudo-amino acid compositions. We achieved 91.24% prediction accuracy in a jackknife test with 49 selected features. It has the best performance among all the present predictors. This result indicates that our computational model can be useful in identifying Golgi-resident protein types.

1. Introduction

In the field of predicting subcellular locations of proteins, a number of models have been developed with different sequence representations and various machine learning methods (Chou, 2009). Many user-friendly online servers were also constructed to increase the availability of the predictive models (Chou and Shen, 2008; Shen et al., 2008; Chou et al., 2011, 2012; Lin et al., 2013; Wu et al., 2011, 2012; Xiao et al., 2011; Du et al., 2011, 2009; Du and Li, 2006).

The Golgi apparatus is an important subcellular organelle in eukaryotic cells. It plays an indispensable role in the secretory pathway. The Golgi apparatus consists of two main parts: the *cis*-Golgi network (CGN) and the *trans*-Golgi network (TGN). The CGN receives the biosynthetic outputs that come from the endoplasmic reticulum. When proteins and lipids are transferred from the CGN to the TGN, modifications are made to these molecules to mark them as functional states. The TGN packages the molecules into membrane-bounded vesicles, and delivers the vesicles to their specific destinations. A number of proteins are retained in the CGN and the TGN to help the above process. We call these proteins the Golgi-resident proteins. We term the proteins that are retained in the CGN and the TGN as the *cis*-Golgi proteins and the *trans*-Golgi proteins, respectively.

As indicated in existing studies, the defects of Golgi apparatus are associated with many kinds of neurodegenerative diseases, including Parkinson's disease (Fujita et al., 2006) and Alzheimer's disease (Gonatas et al., 1998). Correctly identifying the types of Golgi-resident proteins may provide informative hints in understanding the functions of Golgi-resident proteins. However, the experimental approaches toward this information are costly and time-consuming. It is urgent to develop computational methods to

* Corresponding author.
E-mail addresses: jiaoyasen@tju.edu.cn (Y.-S. Jiao), PufengDu@gmail.com (P.-F. Du).

http://dx.doi.org/10.1016/j.jtbi.2016.04.032
0022-5193/

identify Golgi-resident protein types using only the primary sequences.

Several computational methods have been developed for this purpose. Ding et al. (2011) proposed to use pseudo-amino acid compositions to predict Golgi-resident protein types with modified Mahalanobis discriminant algorithm. They improved their methods by using g-gapped dipeptide compositions and feature selection methods (Ding et al., 2013). We proposed to incorporate ANOVA feature selection method to predict Golgi-resident protein types (Jiao and Du, 2016). van Dijk et al. (2008) proposed a method to predict Golgi-resident protein types for type II membrane proteins using structural information and transmembrane domain.

Chou's pseudo amino acid compositions (PseAAC) (Chou, 2001) have been widely applied in predicting a wide range of protein cellular attributes (Chou, 2011). A large number of works have taken the PseAAC as their basis of success (Liu et al., 2015a, 2015b, 2015c, 2015d). Recently, the concept of PseAAC has been extended to represent DNA and RNA sequences (Chen et al., 2013). The Pseudo-k nucleotide compositions (PseKNC), which is a novel nucleotide sequence representation, has been applied in predicting attributes of DNA sequences (Chen et al., 2015). Moreover, various kinds of software resources have been released for converting sequences into their PseAAC or PseKNC representations (Cao et al., 2013; Shen and Chou, 2008; Du et al., 2014, 2012; Chen et al., 2014, 2015; Liu et al., 2015).

On molecular level, the evolutionary process involves insertions, deletions and changes of residues in protein sequences. To incorporate the evolutionary information into PseAAC, Shen and Chou (2007b) proposed the pseudo-positional specific scoring matrix (PsePSSM) method in predicting protein subnuclear locations. However, the PSSM may not exist for some proteins. The PsePSSM method must incorporate an alternative model in case that the PSSM cannot be generated for some proteins. To avoid this trouble, we proposed the positional specific physicochemical properties (PSPCP), which integrate the information of PSSM into artificial physicochemical properties. When the PSSM does not exist for some proteins, the PSPCP representations will automatically degrade to a compatible PseAAC representation without any manual interference. The PSPCP method has been successfully applied in predicting submitochondrial locations (Du and Yu, 2013).

In this paper, we propose to use the PSPCP representations in couple with the minimum Redundancy Maximum Relevance (mRMR) (Ding and Peng, 2005; Peng et al., 2005) feature selection scheme to identify the types of Golgi-resident proteins. Our method performs better than the existing state-of-the-art methods.

2. Materials and methods

2.1. Dataset curations

Ding's dataset (Ding et al., 2013), which was obtained from http://lin.uestc.edu.cn/server/SubGolgi/data, was applied in this work. According to the description of Ding et al. (2013), they extracted this dataset from the UniProt database with the following steps:

(1) Only those proteins, which have been clearly annotated with *cis*-Golgi or *trans*-Golgi, have been collected. The proteins with *cis*-Golgi annotations were categorized as the *cis*-Golgi proteins. The proteins with *trans*-Golgi proteins were categorized as the *trans*-Golgi proteins.
(2) Only those proteins with experimentally verified annotations were kept. All the proteins, which are annotated with 'PROBABLE', 'POTENTIAL', 'POSSIBLE' or 'BY SIMILARITY', were discarded.
(3) The protein sequences with ambiguous amino acid notations (X, B or Z) were discarded as well as the fragments of other proteins.
(4) The sequence similarity level was controlled to be lower than 25%. The sequence length was controlled to be between 25 and 10 Kaa.

After all above screening procedures, a benchmarking dataset, which contains 42 *cis*-Golgi proteins and 95 *trans*-Golgi proteins, was obtained.

2.2. Sequence representations

We applied PSPCP in couple with the PseAAC concept in this work. The PSPCP integrates the PSSM information within artificial physicochemical properties. These artificial physicochemical properties were used to replace the conventional physicochemical properties in the PseAAC. We have elaborated the PSPCP method in (Du and Yu, 2013). For the reader's convenience, we briefly describe this method here as follows.

Let $P=R_1R_2\ldots R_L$ be a protein sequence of length L, where R_1, R_2, …, R_L are L residues on P. We used PSI-BLAST to search P against the UniProt database for three iterations. The e-value threshold was set to 0.001. A PSSM matrix was generated for every P, as follows:

$$E(P)=\begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to20} \\ \vdots & \vdots & \ddots & \vdots \\ E_{L\to1} & E_{L\to2} & \cdots & E_{L\to20} \end{bmatrix}, \tag{1}$$

where $E_{i\to j}$ is a score produced by the PSI-BLAST (Altschul et al., 1997). $E_{i\to j}$ represents the tendency that the i-th residue of P is transferred to the j-th type of amino acid during the evolutionary process. $E_{i\to j}$ varies in a wide range. We normalized each element in $E(P)$ to range [0, 1], as follows:

$$A(P)=\begin{bmatrix} A_{1\to1} & A_{1\to2} & \cdots & A_{1\to20} \\ A_{2\to1} & A_{2\to2} & \cdots & A_{2\to20} \\ \vdots & \vdots & \ddots & \vdots \\ A_{L\to1} & A_{L\to2} & \cdots & A_{L\to20} \end{bmatrix}, \tag{2}$$

where

$$A_{i\to j}=\frac{\exp(E_{i\to j})}{\sum_{j=1}^{20}\exp(E_{i\to j})},\ i=1,\ 2,\ \ldots,L;\ j=1,\ 2,\ \ldots,\ 20. \tag{3}$$

We defined the PSPCP using the values in matrix $A(P)$. Let $H(r,j)$ be the r-th physicochemical property of the j-th type of residue. The r-th type PSPCP for the i-th position of P is given as follows:

$$d_{i,r}(P)=\sum_{j=1}^{20}A_{i\to j}h(r,j), \tag{4}$$

where $d_{i,r}(P)$ is the r-th type PSPCP for the i-th position of P, and $h(r,j)$ the normalized r-th physicochemical property of the j-th type of residue. The $h(r,j)$ was defined as follows:

$$h(r,j)=\frac{H(r,j)-m(r)}{s(r)}, \tag{5}$$

where

$$m(r)=\frac{1}{20}\sum_{j=1}^{20}H(r,j),\ \text{and} \tag{6}$$