Biochemical and Biophysical Research Communications 399 (2010) 402–405

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Prediction of protein subcellular localization by weighted gene ontology terms

Sang-Mun Chi *

School of Computer Science and Engineering, Kyungsung University, 110-1, Daeyeon 3 dong, Nam-gu, Busan 608-736, Republic of Korea

ARTICLE INFO

Article history:
Received 19 July 2010
Available online 3 August 2010

Keywords:
Protein subcellular localization
Text categorization
Term-weighting
Gene ontology
Chi-square

ABSTRACT

We develop a new weighting approach of gene ontology (GO) terms for predicting protein subcellular localization. The weights of individual GO terms, corresponding to their contribution to the prediction algorithm, are determined by the term-weighting methods used in text categorization. We evaluate several term-weighting methods, which are based on inverse document frequency, information gain, gain ratio, odds ratio, and chi-square and its variants. Additionally, we propose a new term-weighting method based on the logarithmic transformation of chi-square. The proposed term-weighting method performs better than other term-weighting methods, and also outperforms state-of-the-art subcellular prediction methods. Our proposed method achieves 98.1%, 99.3%, 98.1%, 98.1%, and 95.9% overall accuracies for the animal BaCelLo independent dataset (IDS), fungal BaCelLo IDS, animal Höglund IDS, fungal Höglund IDS, and PLOC dataset, respectively. Furthermore, the close correlation between high-weighted GO terms and subcellular localizations suggests that our proposed method appropriately weights GO terms according to their relevance to the localizations.

1. Introduction

Knowledge about protein subcellular localization (PSL) provides important information about protein function, because PSL and protein function are highly correlated. Many computational approaches have been developed for PSL prediction, which can be classified according to the information source used in the prediction methods. Widely used information source is sequence-based features, such as amino acid composition and sorting signals [1–7]. Another type of information is textual descriptions of proteins [8–10]. Recently, gene ontology (GO) annotation has been used as the information source for PSL prediction [11–15]. GO annotation is the association of GO terms in three domains, namely, molecular function, biological process, and cellular component, with gene product properties [16]. GO-based prediction methods perform well because GO annotation and PSL are strongly correlated.

Currently, GO-based methods use all of the GO terms in a training dataset as the information source or select a small number of informative GO terms. This study assigns a weight to each GO term according to its discriminative power in order that high-weighted GO terms contribute more to PSL prediction than low-weighted terms. To assign these weights, we employed the term-weighting methods used in text categorization. Text categorization methods classify documents into predefined categories by mapping the contents of documents into a representation which can be interpreted by a machine-learning algorithm [17,18]. This representation is a vector of term weights; term-weighting methods are used to weight individual terms according to their capability to represent the document. Although the term-weighting methods used in text categorization have been compared in extensive literature, this study evaluates the weighing methods in the context of PSL prediction. We also propose a new term-weighting method and compare its performance with existing methods.

2. Methods

To identify informative GO terms for PSL prediction, we investigate several term-weighting methods used in text categorization. In text categorization, a document is classified into predefined categories, $C = \{c_1, \ldots, c_{|C|}\}$. First, a document d_j is represented as a vector of term weights:

$$\vec{d}_j = \langle w_{1j}, \ldots, w_{|T|j} \rangle \quad (1)$$

where the set T contains all the terms (typically, words) occurring in the training documents, $Tr = \{d_1, \ldots, d_{|Tr|}\}$. For the term $t_k \in T$ and document d_j, the weight w_{kj} is the contribution of t_k to the discriminative semantics of d_j. To formulate w_{kj}, the most widely used term-weighting method is *tf idf* function:

$$tf\,idf(t_k, d_j) = tf(t_k, d_j) \cdot idf(t_k) \quad (2)$$

where the text frequency $tf(t_k, d_j)$ denotes the number of times t_k occurs in d_j, and the inverse document frequency $idf(t_k)$ is defined as

$$idf(t_k) = \log(|Tr|/\#Tr(t_k)) \quad (3)$$

* Fax: +82 51 622 1078.
E-mail address: smchiks@ks.ac.kr

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.07.086

where $\#Tr(t_k)$ denotes the number of documents in Tr in which t_k appears at least once. By using $idf(t_k)$, a term occurring in only a few training documents is considered to be a good discriminator of the documents. The values obtained by Eq. (2) are usually normalized using the equation:

$$w_{kj} = tfidf(t_k, d_j) \setminus \sqrt{\sum_{s=1}^{|T|} tfidf(t_s, d_j)^2} \quad (4)$$

The weight w_{kj} can be enhanced by replacing $idf(t_k)$ in Eq. (2) with more effective functions. As a replacing function, we test several term-selection functions used in text categorization; in order to concisely convey the meaning of documents, term-selection methods reduce the size of the set T by selecting only specific and informative terms for each category [17,18]. In the following term-selection functions, all of the probabilities are estimated by counting the number of occurrences of a term in training documents. For example, $P(t_k, \bar{c}_i)$ denotes the probability that a term t_k occurs in a document which does not belong to the category c_i. Information gain (IG) is defined as the amount of information t_k contains about c_i based on the presence or absence of a term in a document [17,18]:

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_i, \bar{t}_i\}} P(t, c) \cdot \log[P(t, c) \setminus (P(t) \cdot P(c))] \quad (5)$$

Since IG grows with the entropy of its variables, dividing IG by the entropy of one variable enables comparison on an equal basis. Gain ratio (GR) is defined as the ratio between IG and the entropy of terms or categories [19]:

$$GR(t_k, c_i) = IG(t_k, c_i) \setminus [-\sum_{c \in \{c_i, \bar{c}_i\}} P(c) \cdot \log(P(c))] \quad (6)$$

Odds ratio (OR) estimates the difference of the distribution of terms in relevant and non-relevant documents [20]:

$$OR(t_k, c_i) = [P(t_k|c_i) \cdot (1 - P(t_k|\bar{c}_i))] \setminus [(1 - P(t_k|c_i)) \cdot P(t_k|\bar{c}_i)] \quad (7)$$

Chi-square (CHI) measures the lack of independence between a term and a document category [17,18]. Specifically, it calculates the difference between the observed and expected frequencies of terms under the assumption of independence. If the difference is large, then the variables are considered to be 'not independent'.

$$\mathrm{CHI}(t_k, c_i) = |Tr| \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]^2 \setminus [P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)] \quad (8)$$

In this study, we evaluate two variations of CHI, the NGL coefficient [21] and the GSS coefficient [22]. The NGL coefficient is defined as

$$\mathrm{NGL}(t_k, c_i) = \sqrt{\mathrm{CHI}(t_k, c_i)} \quad (9)$$

And the GSS coefficient is defined as

$$GSS(t_k, c_i) = P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i) \quad (10)$$

In addition, we proposed a novel logarithmic transformation of CHI, which is defined as

$$LCHI(t_k, c_i) = \log(1 + \mathrm{CHI}(t_k, c_i)) \quad (11)$$

Functions of the form $f(t_k, c_i)$, such as Eq. (5)–(11), can be used to identify terms that are distributed differently between categories. In this study, the difference in the distribution of t_k is represented as a score s_k, which is defined as

$$s_k = \max_{i=1}^{|C|} f(t_k, c_i) \quad (12)$$

or

$$s_k = \sum_{i=1}^{|C|} f(t_k, c_i) \quad (13)$$

In addition to these term-weighting methods, we include no-weighting (NW) method:

$$s_k = 1, \text{for all } k \quad (14)$$

Finally, this score s_k replaces $idf(t_k)$ in Eq. (2) to enhance the discriminative power of w_{kj} in Eq. (4).

For PSL prediction, we apply these term-weighting methods used in text categorization to weight GO terms of proteins. GO annotation is the association of gene product properties with GO terms in three domains: (1) molecular functions of the gene products, (2) biological processes involving the gene products, and (3) subcellular localization of the gene products. For PSL prediction, we make a following correspondence: document, a term (e.g., a word), and a document category in text categorization are mapped to a protein, a GO term, and a PSL, respectively. Proteins are represented with GO terms as the following procedure.

(1) A FASTA format database is prepared. The elements of this database are extracted from Swiss-Prot release 57 [23] (downloaded from http://www.uniprot.org/downloads), which have a GO annotation in a file 'UniProt'. The file 'UniProt' (from UniProtKB-GOA UniProt version 81 [24]) contains associations between gene products and GO terms, it is downloaded from http://www.ebi.ac.uk/GOA.
(2) For each protein, BLAST [25] with default parameters is used to search for a homology in this FASTA format database, and then the GO terms of this homology were retrieved from the file 'UniProt'.

Subsequently, each protein p_j is represented as a vector of term weights as in Eq. (1):

$$\vec{p}_j = \langle w_{1j}, \cdots, w_{|T|j} \rangle \quad (15)$$

where the set T contains all of the GO terms of proteins in the training set $Tr = \{p_1, \cdots, p_{|Tr|}\}$. The term weight w_{kj} is calculated as follows. First, a set $S = \{s_1, \cdots, s_{|T|}\}$ is calculated using one of Eq. (12)–(14), and then the score s_k substitutes for $idf(t_k)$ in Eq. (2). Second, $tf(t_k, d_j)$ in Eq. (2) is defined as 1 if t_k exists in p_j; otherwise, it is defined to be 0. Finally, Eq. (4) is used to normalize the values obtained by Eq. (2) to produce the term weights in Eq. (15).

3. Results and discussion

For the prediction experiments, three benchmark datasets were used: BACEL, HOGL, and PLOC. Two datasets in the BACEL, BaCelLo dataset [4] and the BaCelLo independent dataset (IDS) [26], were used for the training and test sets, respectively. The BaCelLo dataset was extracted from Swiss-Prot release 48 and contains 2597 animal proteins, 1198 fungal proteins, and 491 plant proteins. The BaCelLo IDS was extracted from Swiss-Prot release 54 and consists of proteins with less than 30% sequence identity with proteins from Swiss-Prot release 48. Clustering all protein sequences that have the same localization and less than 30% sequence identity resulted in 432 animal groups, 418 fungal groups, and 132 plant groups. The BACEL covers 4 localizations for animal and fungal proteins (nucleus, cytoplasm, mitochondrion, and secretory pathway) and 5 localizations for plant proteins (chloroplast, in addition to those for animal and fungal proteins).

Similarly, two datasets in the HOGL, the Höglund dataset [5] and the Höglund IDS [13] were used for the training and test sets, respectively. The Höglund dataset was extracted from Swiss-Prot release 42 and contains 5959 eukaryotic proteins. The Höglund IDS was extracted from Swiss-Prot release 55.3 and clustered in the same way as the BaCelLo IDS, resulting in 158 animal groups, 106 fungal groups, and 30 plant groups. The Höglund IDS covers