I J B R
International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 2, 2011, pp-214-220
Available online at http://www.bioinfo.in/contents.php?id=21

# GENE FUNCTION PREDICTION BY THE MULTI-LAYERED CLASSIFIER WITH MULTI-FEATURES

**JAEHEE JUNG, GANGMAN YI***
Samsung Electronics Co., Ltd. 416, Maetan 3-dong, Yeongtong-gu, Suwon-si, Gyeonggi-do, 442-742 Korea
*Corresponding Author: Email- gangman.yi@gmail.com

Received: October 24, 2011; Accepted: November 15, 2011

**Abstract-** Gene Ontology (GO) is a controlled vocabulary to describe the gene function. Each GO term is constructed by a hierarchical structure of gene function, so it is suitable for describing relationships of gene functions. We applied Bayesian network model as a training model using GO terms to identify unknown gene functions, and used Bayesian network model with three different heterogeneous data sets and multi-layered classifier to automatically predict gene functions. This proposed model is comprised of a base-classifier and a meta-classifier. The base-classifier serves a base of meta-classifier with Bayesian network model and meta-classifier plays role of classifying the designated GO term from the root node. A comparative analysis of our suggested model and other gene functional annotation systems shows that our model outperforms than others especially in terms of a number of correctly predicted proteins.
**Key words** - Gene Ontology, DAG, Hierarchical, Gene annotation, meta-classifier

## Introduction

The high-throughput genome sequencing technology has increased and generated tons of unknown sequences, so the development of automated annotation method is becoming increasingly important to biologists in a useful way of identifying the function of genes. Many researches employed various features such as the sequence similarity[1], gene expression[2] or protein domain database[3] to predict the gene function, however, most research did not considered the hierarchical gene information to simplify the classification. To express the hierarchical gene structure, the Gene Ontology [4] is the most suitable mechanism. The Gene Ontology is a database by a controlled vocabulary of terms that describe functions of genes and gene products. GO is constructed by a directly acyclic graphs (DAG) that describes the relationships among similar terms.

Several authors have recently researched about the gene annotation by hierarchical information[5,6,7,8]. Shahbaba and Neal [5] suggested the models by employing the hierarchical structure, however, they consider only one parent in each node. This is not fully supported by attributes of the GO structure, since GO term has at least one parent. Barutcuoglu et al. [6] also used a Bayesian network for the purpose of developing a multi-label annotation method, overcoming the shortcoming of inconstancy between the child and parent annotations by preventing child terms from being annotated. Sokolov and Ben-Hur[7] suggested the hierarchical predicted model using the kernel method for structured output space. Jung et al. [8] report that the hierarchical GO-structured model with protein domain presents. But this strategy has a shortcoming that means each protein has to be one or more one InterPro terms to used as a training data or test data set. This implies that the protein, which does not have a InterPro term, cannot be annotated. To overcome those restrictions, we suggest a model with multi-features, multi-layered classifier and the Bayesian network.

## Methodology

PoGO[9] represents that the performance used many feature lists is better than employing only one features[7]. However, both methods are not considered the GO term's hierarchical information. Therefore, we suggest the new model which is multi-layered with multi-feature by considering GO hierarchical structured data.

### a. Data Set

In the PoGO[9], four feature sets are employed, which are InterPro terms, BLAST, protein structure information and bio-chemical properties. Among these four feature sets, all except bio-chemical information can be converted to the binary format. The bio-chemical property is a numeric value, so this feature sets are excluded in the data. Originally, InterPro terms demonstrate the binary data and BLAST features can be indicated in a true or false format instead of as a cumulative counted value. Protein structure information can also be applied as a binary data from the SCOP database instead of as a score value. The treated data is composed of 8208 proteins and 3339 InterPro terms, 3182 BLAST and 8494 protein structure information. From the GO classifier point of view, similar to the hierarchical GO-structured model[8], all parent GO terms also include the training data sets. Thus, the number of GO terms that can be trained are increased to 4706 from 3182.

**b. Training Procedure**

The multi-layered classifier is made up of base-classifier in each feature set and a meta- classifier. The base-classifier servers to build the meta-data for the training set and the meta- classifier plays role of classifying with heterogeneous data from the base-classifier result. In other word, the multi-layered learning scheme from the meta-classifier enables us to merge the different feature sets, resulting in improving of the performance. Both PoGO[9] and this suggestion use the multi feature sets, but the difference is that this suggestion is a base-classifier learning scheme by the hierarchical model rather than an independent learning SVM with features and instances selection. In addition, the proteins which do not have InterPro terms are also used in the training set. Therefore, the number of annotation GO terms are larger than PoGO and the accuracy is higher than that.

The experiment is performed by 10-fold cross validation. The base-classifier in each feature set is the probabilistic model of Naive bayes. Since the data format is binary and each feature list is independent, the Bayesian probabilities are easily calculated. As in the learning scheme described in hierarchical GO-structured model[8], the training model is composed of a Bayesian probabilistic matrix. Hence, the whole training models are 3339*4706*2, 3182*4706*2, 8492*4706*2 for the InterPro terms, BLAST and protein structure information, respectively. In the InterPro meta-data sets, two subsets from the InterPro training sets are separated randomly. One subset is used for the training and the other set is used for the testing and vice versa, resulting in creating the meta-data. When two subsets are tested independently, the meta-data is assigned true if the true probability is larger than false, hence it is a binary formed column. Given this algorithm, other two more feature sets are also trained. Finally, 7428*3 probabilistic matrices are obtained, where 7428 is the number of training proteins in each validation, because the meta-data set is composed of the test results of the other subset's training model. With this meta-data, the meta- classifier is learned by the Naive bayes. The three binary column sets and the designated GO classifier can be computed the by the Bayesian probability, resulting in it being constituted by 3*4706*2 Bayesian probabilistic matrices for the meta-classifier, where 4706 is the GO term number and 3 is the feature set's meta-data and 2 stands for true and false.

**C. Test Procedure**

The basic formula for the base classifier is the same as GO-structured model with InterPro [8], while, only difference is this suggestion applies three feature sets separately in a base-classifier. Since we have three feature sets, the base classifier in each GO term is represented by $Xv^j$ , where j is one of the InterPro terms, BLAST and protein structure. If $P(X1^j, X2^j ...., M^j_{v=T})$is larger than $P(X1^j, X2^j ...., M^j_{v=F})$ ,the $v^j$is assigned true, otherwise false, this is annotated as $M^j_v$. Therefore, $Mv^j$ is the meta-data executed by the base-classifier. After that, a meta-classifier using this meta-data is accomplished for the purpose of the integration of heterogeneous data. The test process with the meta-classifier is accomplished by $\prod_{j=1}^{3}(G_{v=T,F}|M_v^j)$in terms of GO terms v, since each feature is independent and the treated feature sets are three, where *j* stands for three feature list.

$$P(X_1^j, X_2^j ...., X_{v=T}^j) = \prod_{i=1}^{v} P(X_{j=\in\{T,F\}}^j | Par(X)_{i=\in\{T,F\}}^j)$$

,

where $X_v^j$ are GO terms in the GO structure with the feature sets, *Par* means parent terms and *j* is features - InterPro terms, BLAST and protein structure.

$P(X_v^j) = P(G_v^j | I_1 \cdots I_k) = \frac{P(G_v^j) P(I_1 \cdots I_k | G_v^j)}{z}$ . $I_k$ are feature lists which the tested protein has according to feature *j* such as IPR1,IPR2. Z is the normalized constant value. Above formula can be simplified like $P(I_1 \cdots I_k | G_v^j) = \prod_{i=1}^{k} P(I_i | G_v^j)$ , where I*i* is independent I*j* ($i \neq j$). Given three $Mv^j$ data, the meta-classifier, $\prod_{j=1}^{3}(G_{v=T,F}|M_v^j)$ can be applied. If $\prod_{j=1}^{3}(G_{v=T}|M_v^j)$ is larger than $\prod_{j=1}^{3}(G_{v=F}|M_v^j)$ v is finally assigned as a candidate term.

When we assume that parent node name is $GO_1$ and child node name is GO2. If the parent node is assigned true, GO2 is calculated by $P(GO^I_{1=T}, GO^I_{2=T,F})$ in the InterPro terms. If $P(GO^I_{1=T}, GO^I_{2=T})$ is larger than $P(GO^I_{1=T}, GO^I_{2=F})$ the $M_2^I$ is considered true. $M_2^B$ , $M_2^F$ are the meta-data from the BLAST and protein structure information. These sets are also treated as the same approach which is described for InterPro term. The three decision labels $(M_2^I, M_2^B, M_2^F)$ from the base-classifier for $GO_2$ are tested by the meta-classifier. In the meta-learner, overall true or false probability is calculated by the product of all probability. If $P(GO_{2=T}|M_2^I) * P(GO_{2=T}|M_2^B) * P(GO_{2=T}|M_2^F)$ is larger than $P(GO_{2=F}|M_2^I) * P(GO_{2=F}|M_2^B) * P(GO_{2=F}|M_2^F)$ , then $GO_2$ is considered as a candidate term, otherwise $GO_2$ is discarded.

**Discussion**

In this results section, we compare the performance from two points of views. The first comparison is the hierarchical GO-structured model[8] with only InterPro terms and multiple features, which allows us to figure out the effect of the multi feature sets. This result is very similar to the single feature set learning model[6] and PoGO at the condition of the independent GO term. First of all, the experiment is performed on the shared GO terms. The shared GO terms in both applications are 967. The overall averages in sensitivity, precision are 0.3147, 0.5323 in the model with using only InterPro terms, but multi-feature models are 0.3468, 0.5615. The multi-feature models are slightly better than a individual single learner with InterPro terms.

Fig.1(a) shows the number of GO terms and the number of annotated proteins at each cut-off F-measure value, where the F-measure is calculated by the protein-based approach and cut-off means F-measure value is higher than each range. The training models with the multi-features have more GO terms in each range, especially, in the low cut-off level (Fig.1 (a)). In the 0.2 cut-off value,

the model with InterPro terms has 725 terms, but model with multi-features has 829 terms. Fig. 1 (b) represent the cut-off related number of annotated proteins at each classifier-based approach. The protein-based approach is measured the f-measure in each protein with the suggested model and the classifier-based approach is calculated one for each GO classifiers. Fig. 2 summarizes the average F-measure based on the protein-based approach. For a ranges except 0.4 and 0.5, the overall value is similar or slightly better in model with multi-feature sets.

The second comparison is multiple feature learning model with the hierarchical GO-structured model or simply independent classifier as PoGO. This comparison can be represented the meaning of the hierarchical GO structure in the multi-feature sets. In the previous paper GO-structured model, the hierarchical GO-structured model with InterPro terms outperformed that not considering the GO hierarchical structured training model, thus GO-structured model with multi-features is also outperformed than independent training model (Fig. 3, Fig. 4). Fig. 3 shows the number of GO terms at each cut-off F-measure value. Absolutely, the multi-feature sets have more GO terms in each range, since the hierarchical model has trained many more GO terms. The PoGO is trained to only 444 classifiers, but the hierarchical GO structure with multi-features is trained to all fungi GO terms. Based on these GO terms, the number of annotated proteins is described (Fig. 4 (a)). However, another reason why many GO terms are annotated is that two GO terms (GO:0005515: protein binding, GO:00058209 : cytosol) are annotated in many proteins unlike the PoGO. Among 8208 proteins, 2072 proteins have GO:0005515 and 1650 proteins have GO:0005829, that is, the number of annotated proteins in each range (Fig. 4 (a)) depends on these two terms. If two terms have a high F-measure value in terms of a classifier-based approach, the number of annotated proteins are also effected in a lower cut-off F-measure. In PoGO classifier, the classifier-based F-measure value in GO:0005515 is 0.7166. and that of GO:0005829 is 0.5723. However, PoGO has values of 0.4674 and 0.3903, respectively. If we excluded these two terms, in results, another figure (Fig. 4 (b)) can be obtained.

During this process, the average protein-based F-measure is also compared in each range shown in (Fig. 5(a)), where the used proteins are Fig. 4(a). The average F-measure with the reduced proteins (Fig. 4(b)), which excluded the two highly annotated GO terms, is described in Fig. 5(b). In both comparisons, a GO hierarchical structure with multi-features outperforms than an independent multi-layered training model(PoGO). However, model without hierarchical structure also provides a good enough performance. This stands for if there are enough proteins to build the training set, even though hierarchical GO structure is not embedded, the learning scheme provides outperformed result. To evaluate this statement, we also compared the shared GO terms both in PoGO and in the hierarchical information with multi-features. In most cases, performance matrices in PoGO have outperformed those with an embedding GO structure. The overall average for sensitivity, precision and F-measure in PoGO are 0.2433, 0.6339, 0.3127, while the hierarchical GO with multi-features are 0.2418, 0.3533, 0.2476.

From this result, we know that if GO terms have enough positive proteins, i.e, classifier with many annotated proteins, then a meta-classifier without hierarchical GO structured model provides good performance. However, most of the GO terms in the fungi set are very sparse and rare, thus a model applied with the GO hierarchical structure is more reasonable in order to assign gene functions.

## Conclusion

In the previous study, we trained independent GO classifiers, but a GO terms structure form, which points out that a parent's term has a relationship with the child term, is not considered in the training scheme. Thus, we proposed a new method for assigning GO terms to proteins using multi-features with Bayesian network model. The Bayesian frame is a graph based model demonstrating the Bayesian probabilistic relationship between random variables. Many studies have also used this approach usually in order to integrate heterogeneous data. However, we use the GO structure for GO structural properties by constructing GO structure in each category by the meta-classifier. In the base-classifier, three feature sets are predicted by the Bayesian network with GO hierarchical structure and overall, two probabilities given true or false GO terms are calculated in the meta-leaner. The hierarchical GO-structured model with multi-feature outperforms that with training only InterPro terms. The multi-feature learning model also contributes more GO terms and more annotated proteins than without the hierarchical modeling with multi-feature sets (PoGO). In addition, this approach satisfies the consistency of prediction, i.e., it does not predict only high-level (parent) GO terms nor only deeper-level (child) GO terms. If this consistency does not meet, the predicted function is located in the high-level. However, the overall F-measure in the shared GO terms is less than in PoGO. Given this result, we analyze that if GO terms are annotated in enough proteins, the modeling without the hierarchical structure is also well-fitted for annotation. However, most GO terms are so sparse that the hierarchical GO-structured model is needed for the gene functional annotation

## References


[1] Martin D., Berriman M. and Barton G.J., (2004) *BMC Bioinformatics*, 5(1) ,178.

[2] Pavlidis P., Weston J., Cai J. and Noble W.S. (2002*) Computational Biology*, 9(2), 401-11.

[3] Shahbaba B. and Neal, R.M. (2006) *BMC Bioinformatics* , 7(1), 448.

[4] http://www.geneontology.org/

[5] Barutcuoglu Z., Schapire R.E. and Troyanskaya O.G. (2006) *Bioinformatics*, 22(7), 830–836.

[6] Sokolov A. and Ben-Hur A. (2010) *Journal of Bioinformatics and Computational Biol*gy , 8(2), 357-76.
[7] Jung J. and Thon M.R. (2006) *Lecture Notes In Computer Science*, 4316, 65-77.
[8] Jung J. and Thon M.R. (2008) *19th International Conference on Pattern Recognition*, *Published by IEEE Computer Society in IEEE Xplore USA,* 1-4.
[9] Jung J., Yi G., Sukno S.A. and Thon M.R. (2010) *BMC Bioinformatics*, 11, 215.

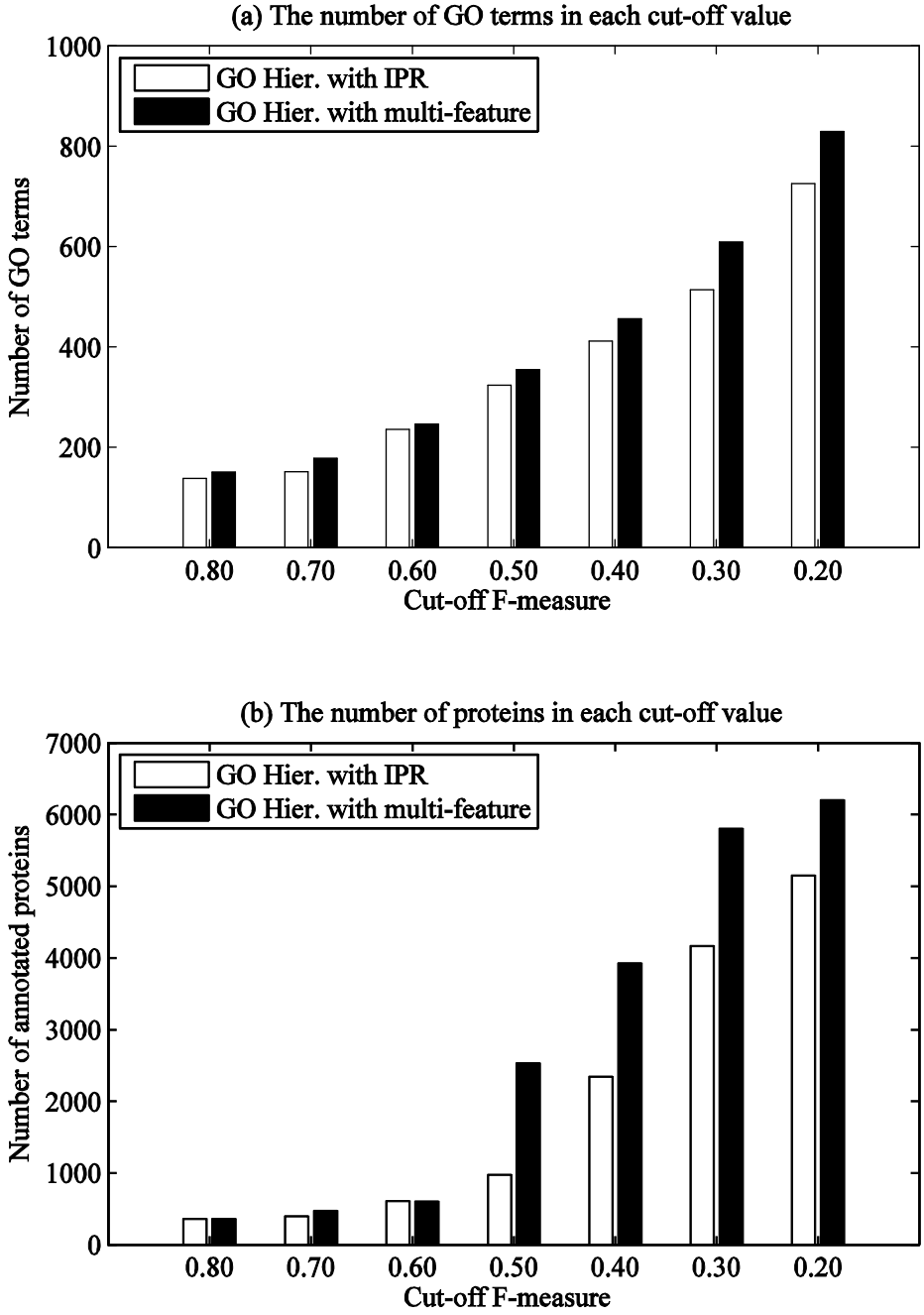


Fig. 1-The hierarchical GO-structured model with InterPro terms or multi-features: (a) The number of GO terms at each cut-o® F-measure value. (b) The number of annotated proteins at each cut-off F-measure value.

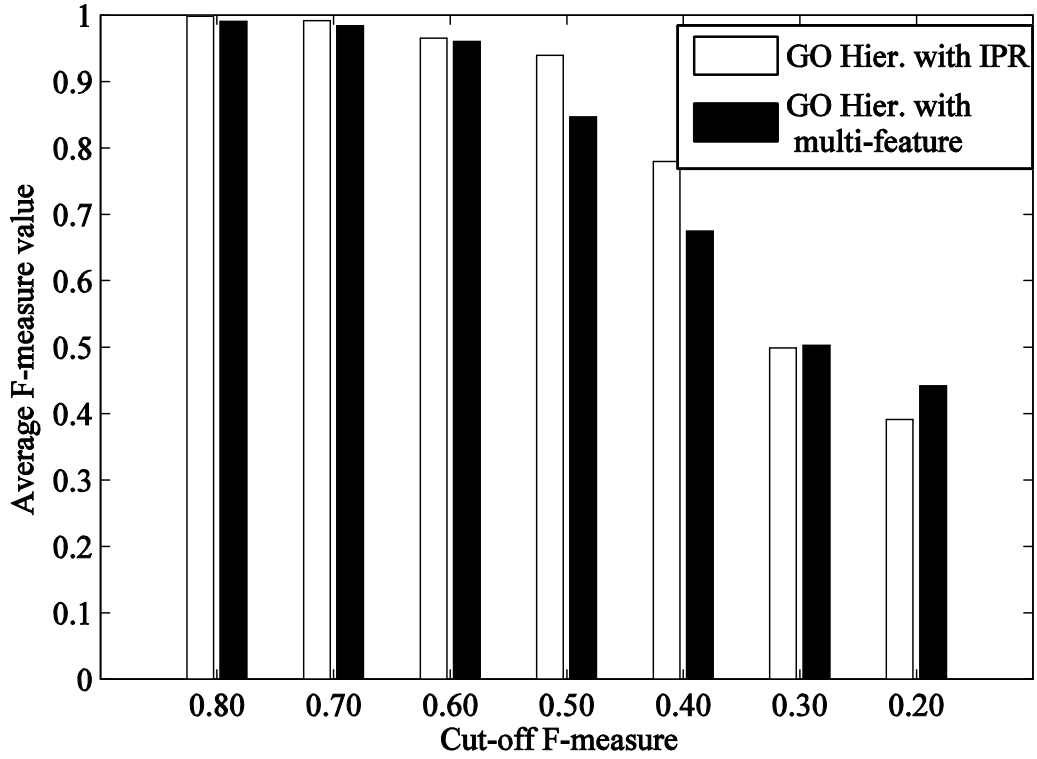


Fig. 2- Average F-measure at each cut-off F-measure value in the hierarchical GO-structured model with InterPro terms and multi-features.

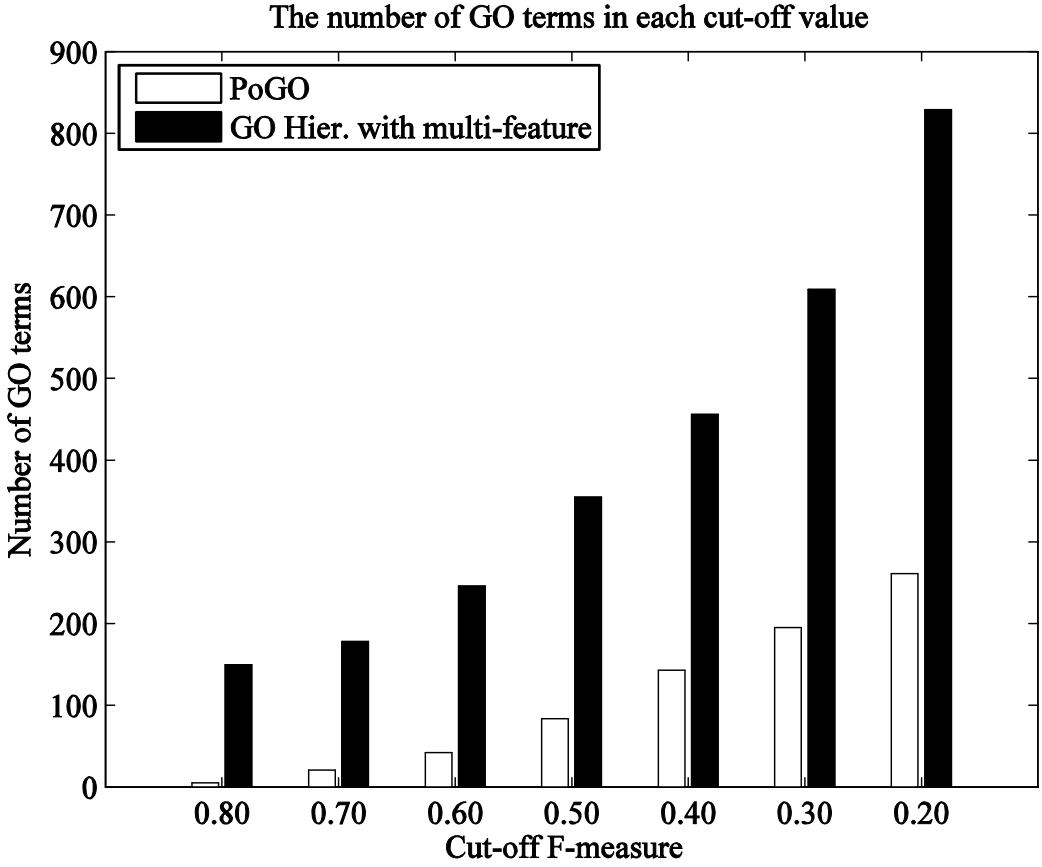


Fig. 3-Number of GO terms at each cut-o® F-measure value in PoGO and the hierarchical GO-structured model with multi-features.

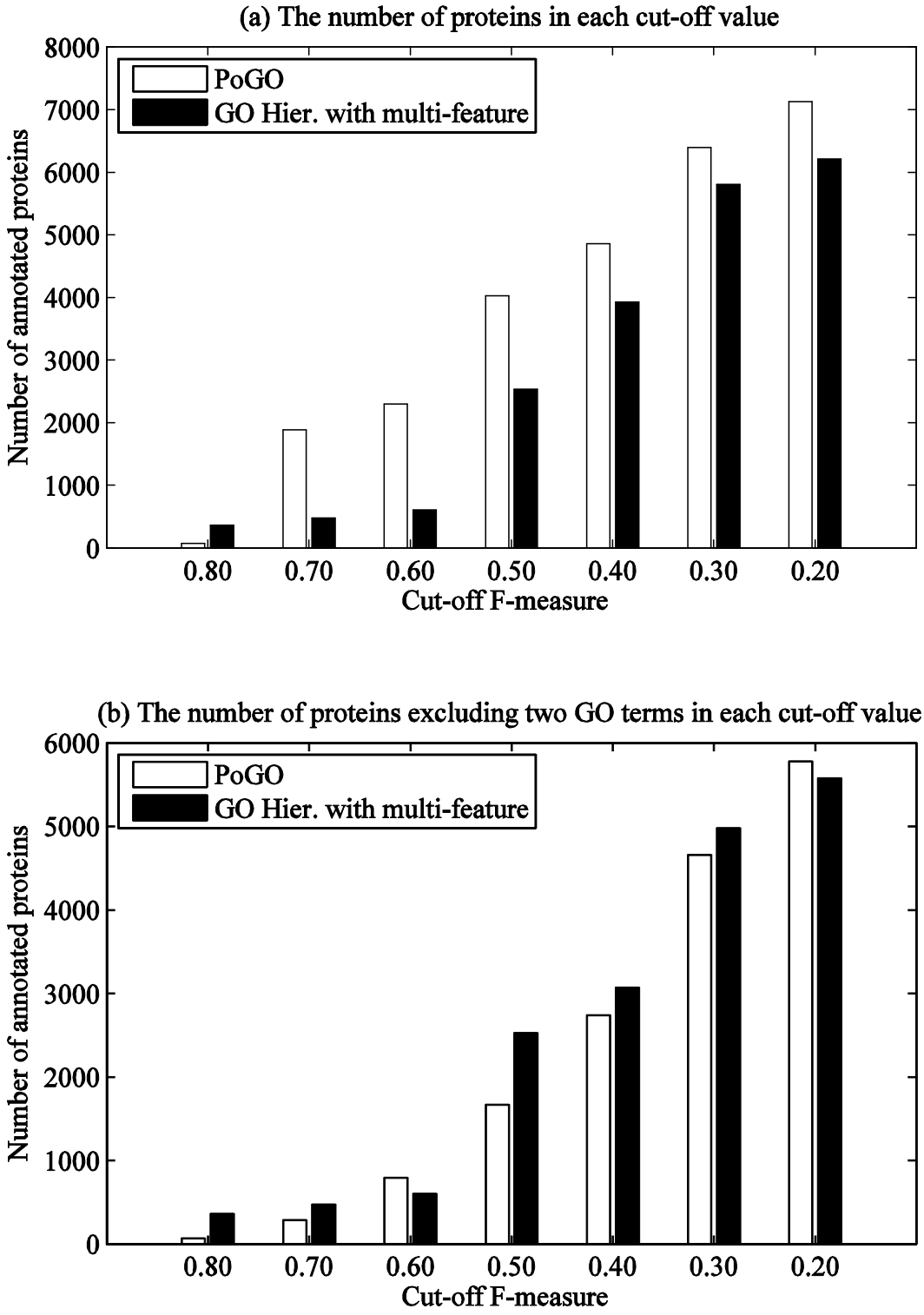


Fig. 4-PoGO and the hierarchical GO-structured model with multi-features set: (a) The number of annotated proteins at each cut-off F-measure value. (b) The number of annotated proteins excluding the two highly annotated GO terms.

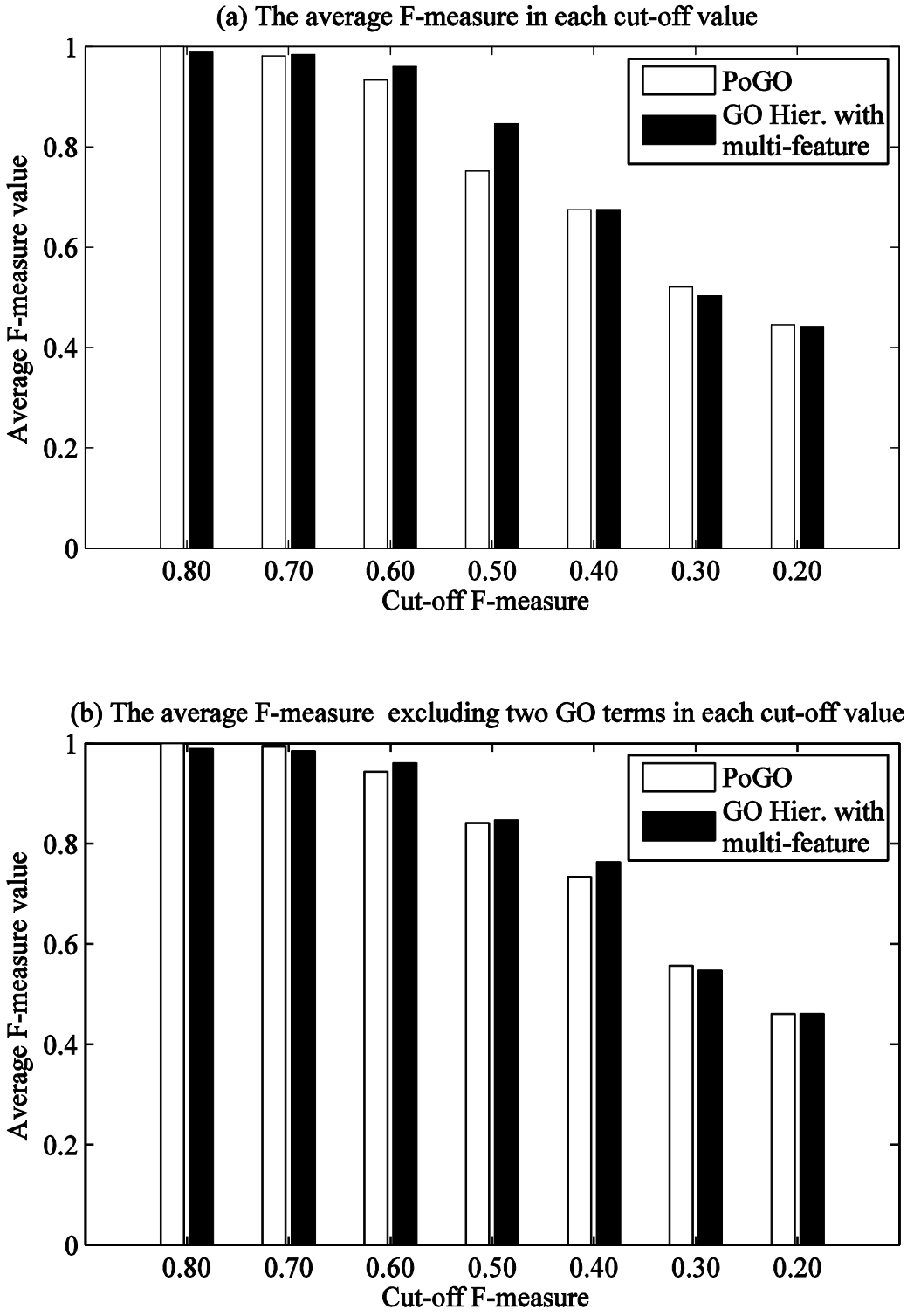


Fig. 5-PoGO and the hierarchical GO-structured model with multi-features set: (a) The average F-measure at each cut-off F-measure value. (b) The average F-measure excluding the two highly annotated GO terms.