A Method to Reveal and Handle Heterogeneities and Inconsistencies in Gene Ontology Annotation

Facchinetti A (1), Sanavia T (1), Di Camillo B (1), Lavezzo E (2), Fontana P (3), Toppo S (4)

(1) Department of Information Engineering, University of Padova, Padova, Italy

(2) Department of Histology, Microbiology and Medical Biotechnologies University of Padova, Padova, Italy

(3) FEM-IASMA Research Center, San Michele all'Adige (Trento), Italy

(4) Department of Biological Chemistry, University of Padova, Padova, Italy

Motivation

Gene Ontology (GO) is a controlled vocabulary of functional terms and is the most widely used annotation database to transfer biological knowledge on gene products. However, it is known that the existing annotations are incomplete and susceptible to several sources of errors (Jones et Al., 2007); in particular, there are gene products whose functions are not completely known/annotated and a large proportion of GO annotations (over 95%) are inferred from electronic annotation (IEA), whereas only a few of them are experimentally validated. In this context, one of the most important challenges is to interpret the accuracy and consistency of the available annotations, in order to identify or correctly predict the functions of protein sequences.

Methods

In this work, we first analyze the global level of heterogeneity in GO annotations to quantify to what extent inconsistencies are present in the GO database. Secondly, we propose a method able to assess annotations of pools of proteins grouped using sequence similarity and to organize the results into a functional map as a useful guidance to easily interpret the reliability of GO annotations. GO annotations were retrieved from GOA database, characterized by both manual and electronic protein annotations within the UniProt Knowledgebase (UniProtKB). Proteins with the same sequence length were grouped using CD-HIT (Li and Godzik, 2006), considering three sequence similarity levels: 100%, 90%, 80%. In order to estimate the proportion of protein groups with possible inconsistencies among the annotations, Quality Threshold clustering (Heyer et Al., 1999), based on similarity between proteins, was applied. A new method was also developed to analyze heterogeneities of GO annotations. For groups of proteins sharing a high sequence similarity, the method performs two agglomerative hierarchical clustering based on semantic similarity between GO terms and protein annotations, respectively. An information-theoretic approach is used to analyze the semantic similar-

ity between GO terms based on the concept of Information Content (Lord et Al., 2003) which represents the ratio between the number of times a GO term and each of its descendants occur in GO annotation, and the frequency of the root term, corresponding to the sum of the frequencies of all GO terms. Semantic similarity is computed using Lin's measure (Lin, 1998) and Best Match Average (Couto et Al., 2007). The algorithm gives as output a colour-coded matrix where each cell (i, j) is coloured if the protein j is annotated with the GO term i, with different colour intensities representing the information content (IC). If the functional map is fully-colored, all the proteins share the same GO terms and the group of proteins is considered as homogeneous with respect to GO annotations. On the other hand, the presence of one or more not-colored cells could highlight possible inconsistencies among GO annotations. In this situation, dendrograms resulting from the two hierarchical clustering on both GO terms and proteins are useful to distinguish different cases of heterogeneities affecting GO annotations. Another important element for the interpretation of the functional map is the colour intensity (i.e. the Information Content), which is highly indicative of the degree of heterogeneity and helps to handle missing or inconsistent biological information.

Results

The global analysis of inconsistencies in the GO database revealed that the percentage of groups of proteins with more than one semantic cluster is around 9.2%, 9.9% and 11.2% for the three sequence similarity levels (100%, 90%, 80%, respectively) in GO Biological Process. Similar values were observed for Molecular Function (5.5%, 6.6%, 7.8%) and Cellular Component (9.9%, 11.0%, 12.1%). These results are highly indicative of the presence of heterogeneities among the GOA annotations and confirm the need of considering the quality and the origin of annotations when inferring a new function based on protein similarity. The developed method facilitates handling this issue. The combined use of Information Content and clustering on both the protein and GO term similarity level is able to efficiently organize information on GO annotations and to highlight in an intuitive and easily interpretable way unexpected and hardly traceable heterogeneities on protein annotations. In particular, the method is able to discriminate missing annotations from inconsistencies and possible errors. The resulting output can be used as a useful guidance to predict functions of new protein sequences or to re-annotate known proteins.

Contact email
stefano.toppo@unipd.it