Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Review

Functional annotation and biological interpretation of proteomics data

Carolina M. Carnielli, Flavia V. Winck, Adriana F. Paes Leme *

Laboratório de Espectrometria de Massas, Laboratório Nacional de Biociências, LNBio, CNPEM, Campinas, Brazil

ARTICLE INFO

Article history:
Received 11 September 2014
Received in revised form 7 October 2014
Accepted 21 October 2014
Available online 31 October 2014

Keywords:
Bioinformatics
Ontology
Annotation
Network
Systems biology
Proteomics

ABSTRACT

Proteomics experiments often generate a vast amount of data. However, the simple identification and quantification of proteins from a cell proteome or subproteome is not sufficient for the full understanding of complex mechanisms occurring in the biological systems. Therefore, the functional annotation analysis of protein datasets using bioinformatics tools is essential for interpreting the results of high-throughput proteomics. Although large-scale proteomics data have rapidly increased, the biological interpretation of these results remains as a challenging task. Here we reviewed basic concepts and different programs that are commonly used in proteomics data functional annotation, emphasizing the main strategies focused in the use of gene ontology annotations. Furthermore, we explored the characteristics of some tools developed for functional annotation analysis, concerning the ease of use and typical caveats on ontology annotations. The utility and variations between different tools were assessed through the comparison of the resulting outputs generated for an example of proteomics dataset.

1. Introduction

Proteomics encompasses a broad range of high-throughput technologies that allows the identification and the quantification of proteins in complex biological samples. Quantitative proteomics approaches rely on the ability to detect small changes in protein abundance of an altered state given a control or reference condition. Thus, the quantification of differences between two or more physiological states of a biological system can be expressed as an absolute protein quantification, by the determination of the exact protein amount or concentration, or as a relative quantification of protein amount, in which the amount of a protein can be defined as fold changes relative to the control sample, determining the up- or down-regulation of such protein [1,2]. Proteomics approaches have been extensively applied in biomedical research for the understanding of diseases, including protein-based biomarker discovery for the early detection and monitoring of different types of cancer [3,4], the analysis of abnormal protein phosphorylation patterns associated with diseases [5,6], such as Alzheimer's [7], the identification of therapeutic targets [8,9], among others. However, mass spectrometry-based proteomics often generates large lists of identified proteins whose interpretation is a challenging task in the field. In order to handle the proteomics data, Biostatistics and Bioinformatics tools become indispensable to the interpretation of biological data and to extract the biological relevance from the vast amount of identified proteins [10]. Thus, protein functional annotation through computational tools now occupies a place as important as the protein identification itself. Since the advent of shotgun proteomics, many Bioinformatics tools have been developed to provide methodologies for functional annotation of proteomics data. Typical approaches for data interpretation for organisms without an annotated genome include mainly the automated protein annotation as a first step in the data analysis workflow. Protein domains, protein family, subcellular localization and biological function are predicted based on sequence similarity searches [11–14].

Once the protein sequences are functionally annotated, several other tools must be applied to the search for functional patterns and overrepresentation of biological functions or processes in a protein dataset from qualitative or quantitative proteomics data. Further steps in the analysis usually include pathway analysis and the prediction of interaction networks, which are generated through integration of different biological layers of information, such as gene expression and co-expression patterns, protein–protein interactions and protein expression data. Moreover, visualization tools largely contribute to localize the presence of targeted proteins within cellular biological pathways, signaling cascades and metabolic pathways being the most represented ones in proteomic studies.

A variety of commercial and open-source bioinformatics tools for the analysis of proteomics data and statistical tests have been developed. However, with the increased amount of proteomics data new challenges in data handling, analysis and visualization push forward the development of the field of computational proteomics. In order to give an overview of tools and approaches currently applied in proteomics functional annotation, we reviewed and discussed different approaches,

* Corresponding author at: Laboratório Nacional de Biociências (LNBio), Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), 13083970 Campinas, Brazil. Tel.: +55 19 3512 1118; fax: +55 19 3512 1006.

E-mail addresses: carolina.carnielli@lnbio.cnpem.br (C.M. Carnielli), flavia.winck@lnbio.cnpem.br (F.V. Winck), adriana.paesleme@lnbio.cnpem.br (A.F. Paes Leme).

http://dx.doi.org/10.1016/j.bbapap.2014.10.019
1570-9639/© 2014 Elsevier B.V. All rights reserved.

computational programs and strategies recently applied for data interpretation, and how different aspects of the analysis can modify the outcome of proteomics studies.

2. Biological meaning of large proteomics datasets through gene ontology-based annotation approaches

The prediction of the functional role of identified proteins in a biological event involves a first step of gathering information, a task that must be performed before the actual biological data interpretation is achieved and may include genome and proteome annotations. Many tools have been developed to mine several databases of biological information to finally predict a protein function based on sequence similarities. Detailed strategies on genomics and proteomics sequence annotation can be found in previous publications [11–17].

Nevertheless, once the genome and proteome are annotated, one of the most disseminated strategies of proteomics data functional annotation includes the use of ontologies, which can be understood as an explicit specification of a conceptualization [18]. Usually, ontologies are designed with hierarchical classes, communicating definitions with clarity and objectivity, however, keeping extendibility. In Biology, the ontologies for genes and proteins usually describe the classification of the molecules according to their role in the biological systems, using controlled vocabulary, which permits the analysis of relationships between the ontology terms through data integration, retrieval and functional annotation of large datasets [19–22].

In this scenario, Ashburner et al. developed a controlled vocabulary applicable to all eukaryotes, generating the Gene Ontology (GO) Consortium [23], with the aim to overcome the lack of interoperability of genomic databases resultant of the divergent nomenclature of genes and proteins. Every gene or protein can thus be described by a finite number of vocabulary terms, which are classified into one of the three GO-categories or domains: biological process, molecular function or cellular component [23].

It is noteworthy that GO annotations to a term are included in a hierarchy of terms, having a more general annotation at the highest levels of the hierarchy and more specific annotation at lower levels of the hierarchy. Moreover, a GO term of lower hierarchical level (Child term) can have a relation to one or more terms of higher hierarchical levels (Parent term), which can be traced up to one or more of the GO root terms which correspond to the three GO domains (biological process, cellular component or molecular Function). For instance, if a gene is found to be related to ‘actin filament bundle organization’ according to its GO annotation, it will be annotated downwards within the hierarchy of its parent terms, which include ‘actin filament organization’, ‘cytoskeleton organization’ and ‘organelle organization’ (Fig. 1).

Thus, more information can be retrieved from parent terms, which increases the knowledge when making inferences about gene function. On the other hand, researches must consider that GO annotations can be redundant, i.e., a term can be associated to one gene or gene product by more than one annotation. In a recent study, Gillis and Pavlidis [24] found that GO annotations are stable over short periods of time, with losses of semantic similarity for 3% of the genes annotated between monthly GO editions. Thus, some undergo changes in their ‘functional identity’ over time as a result of annotation updates, resulting in loss of semantic similarity matching. Additionally, they presented a way to quantify the stability of GO annotations over time and showed that, in a moderate time, many genes undergo changes in their annotated functionality. Thus, modifications on gene ontologies may influence the results on functional annotation of experimental data [25]. Despite that changes in GO annotations are non-uniformly distributed over different branches of the ontology, the results of term-enrichment analyses were found stable [25].

In order to observe and to demonstrate how different versions of the GO annotations may affect the final interpretation of a proteomics dataset we performed a comparison of the results of the enrichment analysis of GO terms using the application BiNGO v.3.0.2 [26] app in the Cytoscape v.3.0.1 [27] to an example proteomics dataset previously published [8]. We used equal parameters for the data analysis and changes in the significant overrepresented terms were evaluated by comparing the list of the Top 10 most significant overrepresented GO terms.

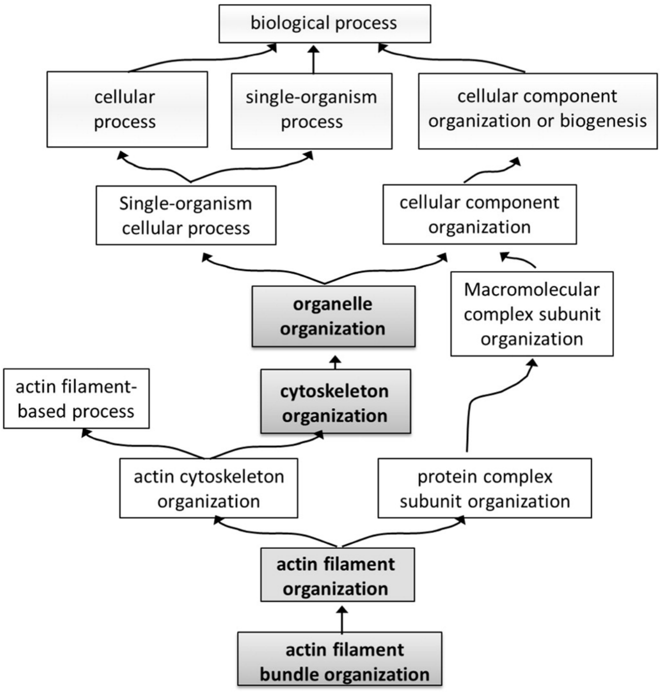
biological process
cellular process
single-organism process
cellular component organization or biogenesis
Single-organism cellular process
cellular component organization
organelle organization
Macromolecular complex subunit organization
actin filament-based process
cytoskeleton organization
actin cytoskeleton organization
protein complex subunit organization
actin filament organization
actin filament bundle organization

Fig. 1. Interconnection of relationships of the hierarchical distribution of GO terms. An example of the hierarchical organization of GO annotations is shown for the GO term “actin filament bundle organization” together with the relationships and intermediate GO terms between the most ancestor (Biological process) to the most specific child GO term (actin filament bundle organization) (Figure adapted from QuickGO — http://www.ebi.ac.uk/QuickGO).

It was observed that changes in the list of the Top 10 most significant GO terms retrieved using GO annotation files from 2011, 2012, 2013 and 2014 occurred with the different annotation files (Supplemental Table 1). However, 60% of the GO terms consistently appeared in the Top10 list of the most significant GO terms, implying a data drift among versions of GO annotation. Nevertheless, most of the GO annotations remain partially stable over time. Thus, it is imperative to perform functional annotation analysis with the most recent version of the GO annotation and ontology files. Moreover, it is important that novel approaches on functional annotation are integrated into dynamic data analysis, allowing on-time updating of annotation files to facilitate and improve the interpretation of published proteomics datasets. Detailed description of parameters and GO association and gene ontology files used in this comparative analysis are available in the Supplemental Table 1.

Furthermore, knowing what has been modified between different versions of the ontology can be very useful. The web service CODEX (Complex Ontology Diff Explorer) was developed to allow users to verify which changes were performed in a precise version of the ontology [28]. However, it is crucial to report, in an ontology-based study, which version of the gene annotation was used in order to track alterations on functional annotation due to time dependence of GO results.

GO annotations can be applied to perform a functional profiling of processes which might be different in a particular set of genes, to predict gene function or to categorize genes in ontology terms [29]. Therefore, the identification of overrepresented categories or enrichment analysis can be performed based on ontologies, contributing to functionally