GOAnno: GO annotation based on multiple alignment

Chalmel F.1‡, Lardenois A.1, Thompson J.D.1, Muller J.1, Sahel J.-A.2, Léveillard T.2, and Poch O.1

1 Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch cedex, France.

2 Laboratoire de Physiopathologie Cellulaire et Moléculaire de la Rétine, Inserm U592, Université Pierre et Marie Curie, 75571 Paris, France.

‡: Corresponding author: chalmel@igbmc.u-strasbg.fr.
ABSTRACT

Summary: GOAnno is a web tool that automatically annotates proteins according to the Gene Ontology (GO) using evolutionary information available in hierarchised multiple alignments. GO terms present in the aligned functional subfamily can be cross-validated and propagated to obtain highly reliable predicted GO annotation based on the GOAnno algorithm.

Availability: The web tool and a reduced version for local installation are freely available at: http://igbmc.u-strasbg.fr/GOAnno/GOAnno.html.

Contact: chalmel@igbmc.u-strasbg.fr

Supplementary information: The web site supplies a detailed explanation and illustration of the algorithm at: http://igbmc.u-strasbg.fr/GOAnno/GOAnnoHelp.html.

INTRODUCTION

Recent efforts in high-throughput sequencing have given rise to a rapid increase in the amount of sequences available in the public databases. Since GeneQuiz (Andrade et al., 1999) that automatically annotated protein function, the systematic annotation of this data is now typically based on the Gene Ontology (GO) (Gene Ontology Consortium, 2000), a hierarchical and standardized vocabulary developed by the GO Consortium (www.geneontology.org). Several tools employ sequence similarities by best BLAST (Altschul et al., 1997) hits selection [e.g. (Hennig et al., 2003; Khan et al., 2003; Zehetner, 2003)] or a predefined subset of GO terms [e.g. (Jensen et al., 2003)].

GOAnno is a web tool for automated protein GO annotation. In contrast to the above methods, GOAnno takes advantage of the evolutionary information available in Multiple Alignments of Complete Sequences (MACS) (Lecompte et al., 2001) organized hierarchically.
into functional subfamilies. The members within subfamilies are conserved enough to filter, enrich and propagate GO terms using the GOAnno algorithm. Another originality is the absence of any predefined parameters such as GO level or subsets of GO terms. The tool uses a query protein sequence as input, and proposes detailed GO annotations in an interactive HTML file as output.

PROGRAM OVERVIEW

The GOAnno algorithm is explained and illustrated in detail in supplementary information. GOAnno incorporates a five step process, numbered from 1 to 5 below.

1 – The query protein functional subfamily determination step incorporates the strategy used in PipeAlign (Plewniak et al., 2003), a toolkit for protein family analysis using a query sequence to perform a protein database sequence search and resulting in a hierarchised MACS of protein homologues clustered into potential functional subfamilies (http://igbmc.ustrasbg.fr/PipeAlign).

The next four steps are independently applied for each of the three GO categories: cellular component, molecular function and biological process. At the end of each step, the redundant and parent GO terms are systematically removed.

2 – An Initial Protein gene Ontology (IPO) is constructed for each query subfamily member from the GO annotation associated with the protein in the sequence databases when available and extracted from the conversion tables available from the GO Consortium (InterPro, Pfam, Prints, PRODOM, Prosite, SMART protein motifs, Enzyme Commission numbers and SWISS-PROT keywords to GO terms).
3 - The construction of the MACS permits the identification of the Proximal Proteins (proteins sharing at least 98 percent identity with the input protein). All the IPO of these proximal proteins are concatenated to form the Proximal Protein gene Ontology (PPO).

4 – The quality of the query subfamily alignment is assessed using the objective scoring function norMD (Thompson et al., 2001). NorMD > 0.3 implies a high-quality and allows the propagation of GO terms within the subfamily according to the following criteria. Briefly, all IPO of the proteins are collected to build the corresponding GO tree. For each IPO, all paths to the root are decomposed into linear branches. Then, a score based on the number of protein is calculated for each node and each branch. Afterward, too specialized nodes and branches associated with rare nodes are eliminated based on two cut-off values $p$ and $f$ respectively. GO terms which pass these selections define the Mean Subfamily gene Ontology (MSO).

5 – The previously determined IPO of the query, PPO and MSO are collected to define the final GPO (Global Protein gene Ontology) that is finally assigned to the query.

In the context of the study of mechanisms leading to retinal degeneration, GOAnno was used on microarray experiments to analyze 1046 UniProt (Apweiler et al., 2004) proteins (Chalmel et al., in preparation). Of these 1046 proteins, 698 had an IPO, corresponding to a total of 2285 GO terms. Using the GOAnno algorithm, GPO were assigned to 191 supplementary proteins (27.4%), corresponding to 1520 new associated GO terms (66.5%).

The interface of GOAnno is designed to accept a single protein sequence as input (Fig. 1). The user has the opportunity to modify the GOAnno parameters (e.g. $f$ and $p$). The program proposes as output a downloadable XML file and an interactive HTML page containing a detailed table describing the IPO, PPO, MSO and GPO steps, where each GO term is linked to
the AmiGO entry (http://www.godatabase.org/cgi-bin/amigo/go.cgi) and each protein accession number to the corresponding UniProt entry.

A light version of GOAnno excluding the first step is also available for local use. In this case, the homologies in terms of subfamily and proximal proteins of the query entry must be previously determined. The program allows automatic batch processing of a gene list, which is of particular interest in interpreting high-throughput experiments such as microarray transcription profiling.

GOAnno provides an efficient way to assign a potential GO to an unknown sequence and to increase an existing GO annotation. It can also be used for in-depth comparisons of functionality relative to a subfamily. GOAnno is designed to help biologists by automatically providing reliable protein functional information combined with an intuitive user interface that can be operated without any previous experience in judging the quality of predicted GO annotation.

ACKNOWLEDGEMENTS

The authors thank G. Berthommier, L. Bianchetti, F. Plewniak, W. Raffelsberger and R. Ripp for stimulating discussions. This work was funded by the INSERM, the CNRS, the ULP de Strasbourg, the INS (GENOPOLE), the SPINE (E.C. contract number QLG2-CT-2002-00988) and the RETNET (E.C. contract number MRTN-CT-2003-504003) projects.
REFERENCES


Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., Prigent, V., Ripp, R., Thierry, J.C., Thompson, J.D.,
