BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 20 2008, pages 2393–2394
doi:10.1093/bioinformatics/btn394

Sequence analysis

Protein function prediction and annotation in an integrated environment powered by web services (AFAWE)

Anika Jöcker, Fabian Hoffmann, Andreas Groscurth and Heiko Schoof*

Plant Computational Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

Received on April 28, 2008; revised on July 16, 2008; accepted on July 26, 2008

Advance Access publication August 12, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Many sequenced genes are mainly annotated through automatic transfer of annotation from similar sequences. Manual comparison of results or intermediate results from different tools can help avoid wrong annotations and give hints to the function of a gene even if none of the automated tools could return any result.

AFAWE simplifies the task of manual functional annotation by running different tools and workflows for automatic function prediction and displaying the results in a way that facilitates comparison. Because all programs are executed as web services, AFAWE is easily extensible and can directly query primary databases, thereby always using the most up-to-date data sources. Visual filters help to distinguish trustworthy results from non-significant results. Furthermore, an interface to add detailed manual annotation to each gene is provided, which can be displayed to other users.

Availability: AFAWE is available at http://bioinfo.mpiz-koeln.mpg.de/afawe/

Contact: afawe-admin@mpiz-koeln.mpg.de

Supplementary information: SIFTER pipeline (S1), AFAWE tutorial (S2).

1 INTRODUCTION

Transfer of functional annotations by sequence similarity from a homolog to a query sequence is the most common method for gene function prediction. However, this general method has many drawbacks (Gilks *et al*., 2002). For further improvement of the functional annotation, one can use several other methods for automatic protein function prediction, for example, structure prediction and comparison, protein domain finding and phylogenomic analysis. Although tools in these fields have increased accuracy, limitations remain and at the moment no tool performs equally well for all kinds of genes. A major problem is the propagation of annotation errors in public databases. Even if new (experimental) function information becomes available, these are rarely updated.

However, by manual comparison of results from different tools, wrong functional annotations can be avoided and the function of a gene can be further specified (Thibaud-Nissen *et al*., 2007). Another advantage is that in some cases the automatic functional prediction of a single tool alone is not able to give a significant result. In this case, the combination of individually insignificant data could give clues to the function of a protein. Unfortunately, this comparison is time consuming, because each tool has its own scores and cutoffs and the user has to switch between different webpages to compare the results. For some tools, there is no web interface available and it is hard to find out how to use these programs.

In the last few years, more and more web services and workflows for biological data retrieval and analysis were published. They improve flexibility by making it easier to include new functionality or data sources; they remove the need for updating local copies of data by working directly on the primary resource; they allow computing-intensive tasks to be executed remotely and improve execution times by parallelizing across distributed resources.

Here, we introduce AFAWE, a tool for *A*utomatic and manual *F*unctional *A*nnotation in a *W*eb services *E*nvironment. It uses web services and Taverna workflows (Oinn *et al*., 2004) to run different function prediction tools. It is easily extensible and wherever appropriate web services are available runs directly on source databases, thus always providing up-to-date data. All results are displayed in an interactive web interface both in graphical and tabular form with extensive filtering options so that trustworthy results are highlighted.

2 METHODS

At the moment, AFAWE includes analyses for homolog detection, protein domain search and phylogenomics. The homolog detection is done by running WU-Blast (Labarga *et al*., 2007) against the UniProt database and separately against the SwissProt database to get a smaller set of homologs with reliable functional annotation. Protein domains are discovered by InterProScan (Labarga *et al*., 2007) and by RPSBlast against the Conserved Domain Database (Marchler-Bauer *et al*., 2003).

For the phylogenomic part, we have implemented a Taverna workflow (see Supplementary Data S1 and http://www.myexperiment.org/workflows/95), which uses SIFTER (Engelhardt *et al*., 2005) to transfer Gene Ontology (The Gene Ontology Consortium, 2000) terms inside the phylogenetic tree by considering duplication and speciation events.

3 WEB INTERFACE

The AFAWE web interface can be entered either by starting a new analysis or by retrieving previous results. A protein sequence and its source organism are the input for a new analysis. Cached results

*To whom correspondence should be addressed.

are searchable by internal and common public database identifiers as well as free text.

The user can select the analysis tools to run and all selected web services and workflows are called in parallel. The results are parsed, stored in a cache database and displayed in several dynamically updated panes accessible through tabs. The graphical and tabular displays aim to allow both a quick overview and thorough browsing of results. The user can come back and view the cached results without running the analyses again. The cached results are deleted if newer results become available (e.g. if databases are updated).

To enable a faster comparison of the results, filters are provided to highlight the most significant analysis results. For the BLAST search, we provide five filters, one to show proteins having the same domain composition as the query protein, and one to highlight proteins overlapping with the query sequence for >70% of their length. The other three filters show hit proteins, which have an experimentally verified or reviewed GO term assigned in one of the three main GO categories: Molecular Function, Biological Process and Cellular Component. GO term assignments are retrieved from UniProt and Gene Ontology annotation files.

We will use the *Medicago truncatula* gene AC144389_35.2 as an example (see Supplementary Data S2). Keyword search for AC144389_35.2 shows four different analysis results (BLAST against UniProt and SwissProt, InterProScan and SIFTER). The SIFTER results are generally more reliable in comparison to BLAST, because they take evolutionary relationships into account. In the 'SIFTER' tab, three different GO terms are shown ['electron transporter, transferring electrons within CoQH2-cytochrome c reductase complex activity' (GO:0045153), 'stearoyl-CoA 9-desaturase' (GO:0004768) and 'enzyme activator activity' (GO:0008047)]. GO term GO:0045153, which has an assigned probability of 0.98, is highlighted as reliable.

By using the GO term filter for Molecular Function, experimentally verified or reviewed GO terms assigned to BLAST hits can be reviewed. Only two proteins (CYB5_YEAST and CYB5_HUMAN) are highlighted, which means that at least one of their assigned functions is verified. Both proteins have >70% overlap with the query and share the same domains with the query (both hits are highlighted in yellow if the overlap and domain filter is switched on) and therefore seem to belong to the same protein family.

The molecular function GO term assigned to *Saccharomyces cerevisiae* gene CYB5_YEAST ['electron carrier activity' (GO:009055)] is experimentally verified by direct assay (evidence code 'IDA') and is the parent of GO:0045153, the term predicted by the SIFTER pipeline. By looking at protein domains predicted by InterProScan, only domains included in Cytochrome b are shown and most BLAST hits are also annotated as Cytochrome b. Cytochrome b is the main subunit of the transmembrane cytochrome bc1 and b6f complexes and is responsible for the transmembrane electron transfer (Howell, 1989). This fits well with GO:0045153.

On the other hand, *Homo sapiens* gene CYB5_HUMAN has GO term 'cytochrome-c oxidase activity' (GO:0004129) assigned by 'traceable author statement (TAS),' and annotation transfer by similarity would pick this up. However, no further support for Cytochrome c oxidase activity can be found; all other sequence and domain matches only support a role as Cytochrome b. Therefore, we assume that this GO term is a wrong annotation and mark it as such in our manual annotation, at the same time marking GO:0045153 as confirmed.

Running a complete automatic annotation with AC144389_35.2 as query, takes about 5 min on our machines, if all found domains and homologous proteins are in the AFAWE database.

4 OUTLOOK

We will use AFAWE in the annotation pipeline of both the international *Medicago truncatula* genome annotation project IMGAG and the international tomato genome annotation project ITAG. In both projects, AFAWE will be used for community annotation. We will extend AFAWE by further methods and analysis results will be combined in a summary page. Web services will be provided to retrieve data from the AFAWE database or to calculate proteins on the fly without using the web interface.

ACKNOWLEDGEMENTS

We thank Barbara Engelhardt for providing us with the SIFTER code and all providers of publicly accessible bioinformatics web services.

Funding: European Commission 6th framework program, project EU-SOL (FOOD-CT-2006-016214).

Conflict of Interest: none declared.

REFERENCES

Engelhardt,B.E. *et al*. (2005) Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.*, **1**, e45.

Gilks,W.R. *et al*. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Howell,N. (1989) Evolutionary conservation of protein regions in the protonmotive cytochrome b and their possible roles in redox catalysis. *J. Mol. Evol.*, **29**, 157–169.

Labarga,A. *et al*. (2007) Web Services at the European Bioinformatics Institute. *Nucleic Acids Res.*, **35**, W6–W11.

Marchler-Bauer,A. *et al*. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*., **31**, 383–387.

Oinn,T. *et al*. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.

Thibaud-Nissen,F. *et al*. (2007) EuCAP, a Eukaryotic Community Annotation Package, and its application to the rice genome. *BMC Genomics*, **8**, 388.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.