ANNOTATING UniProt METAGENOMIC AND ENVIRONMENTAL SEQUENCES IN UniMES

Samuel Patient and Maria Martin
EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, U.K.

Keywords: UniMES, UniProt, Automatic Annotation.

Abstract: This short paper outlines the mechanisms for the system developed in UniProt for function annotations of UniProt Metagenomic and Environmental Sequences in UniMES.

1 INTRODUCTION

The automatic annotation system developed in UniProt for functional annotation of uncharacterised proteins has been adapted to annotate the UniProt Metagenomic and Environmental sequences in UniMES. As a single annotation item (i.e. protein name, enzyme nomenclature, protein function or ontology relationships) can be the result of different sources for annotation and/or related annotation items can be the result of the same source of annotation, the derived evidence is provided through a methodological system of evidence 'tags', where the description of the annotation procedure is provided. This system of evidence tags is essential for both database maintenance and update and for user interpretation of annotations in the resource.

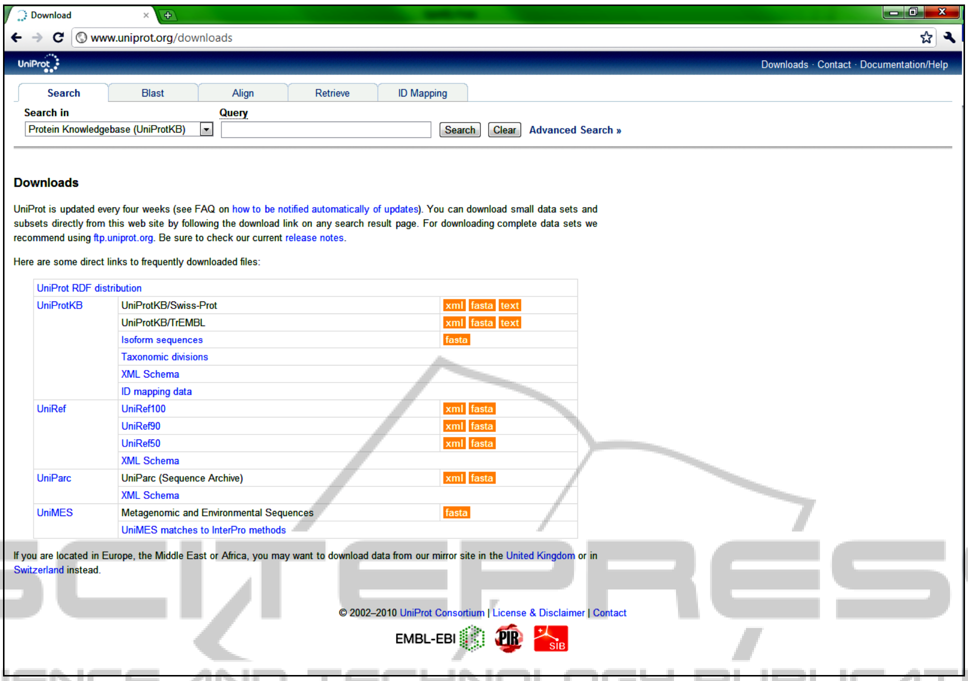
The automatic annotation system developed in UniProt is based upon rules derived from the combination of (1) the protein family classification provided by InterPro and (2) published experimental data incorporated in the fully manually annotated section of UniProtKB (UniProtKB/Swiss-Prot). The use of protein family and domain classifications allows the identification of proteins that are difficult to characterise based on pairwise alignments. It also provides an effective means to retrieve relevant biological information from vast amounts of data as well as reflecting underlying gene families, an analysis of which is essential for subsequence comparative genomics and phylogenetics analysis. This rule-based automated annotation procedure leads to automatic functional annotation of general protein properties which can be conservatively predicted to avoid biological errors, such as function(s) of the protein, domains and sites, catalytic activity, pathways, and subcellular location, and to position-specific information like active sites.

The automated annotation process consists of three steps; a data classification step provides clusters of protein sequences belonging to certain families or domain types; a data mining step then provides annotation rules generated by computational methods and reviewed by expert biologists; finally, a third step covers the application of the output of the data mining procedures onto records of uncharacterised proteins.

(1) The classification step consists of the use of InterPro as an external database to cluster sequences in the high-quality annotated UniProtKB/Swiss-Prot database into groups. InterPro integrates predictive models or signatures representing protein domains, families and functional sites from different member databases such as Pfam, TIGRFAMs and PROSITE (Hunter *et al.,* 2009).

(2) All UniProtKB/Swiss-Prot proteins belonging to an InterPro family or domain type are inspected to identify functional information shared by most of the proteins. In all statistically relevant cases (over 99% of the sequences in the set have common signatures and annotations), an annotation rule is derived and manually inspected for validation. The rules generated consist of the InterPro conditions for application of the rule and the appropriated annotations that result from the conditions being satisfied. Additionally, some rules also provide taxonomic information when this is relevant to a particular set of annotations. Standard data mining algorithms can be used

Patient S. and Martin M..
ANNOTATING UniProt METAGENOMIC AND ENVIRONMENTAL SEQUENCES IN UniMES.
DOI: 10.5220/0003350803670368
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (Meta-2011), pages 367-368
ISBN: 978-989-8425-36-2



Downloads

UniProt is updated every four weeks (see FAQ on how to be notified automatically of updates). You can download small data sets and subsets directly from this web site by following the download link on any search result page. For downloading complete data sets we recommend using ftp.uniprot.org. Be sure to check our current release notes.

Here are some direct links to frequently downloaded files:

UniProt RDF distribution		
UniProtKB	UniProtKB/Swiss-Prot	xml fasta text
	UniProtKB/TrEMBL	xml fasta text
	Isoform sequences	fasta
	Taxonomic divisions	
	XML Schema	
	ID mapping data	
UniRef	UniRef100	xml fasta
	UniRef90	xml fasta
	UniRef50	xml fasta
	XML Schema	
UniParc	UniParc (Sequence Archive)	xml fasta
	XML Schema	
UniMES	Metagenomic and Environmental Sequences	fasta
	UniMES matches to InterPro methods	

If you are located in Europe, the Middle East or Africa, you may want to download data from our mirror site in the United Kingdom or in Switzerland instead.

© 2002–2010 UniProt Consortium | License & Disclaimer | Contact

Figure 1: The UniProt website provides adownload of the UniMES data and its associated InterPro matches.

to detect data dependencies and produce annotation rules automatically. UniProt has made used of the C4.5 decision tree algorithm to represent automatically the proteins of a given training set (UniProtKB/Swiss-Prot in this case) into positive and negative examples depending on the presence of a particular annotation. The system is supported by statistics that support and sometimes replace manual biological checks.

(3) The application step provides the means to transfer the corresponding annotations to the sequence data in UniMES by means of using the matching InterPro signatures (Fig. 1 shows where this data can be retrieved from).

REFERENCES

The UniProt Consortium (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 36, D190–195.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A., Orchard, S., Orengo, C., Petryszak, R., Selengut, J., Sigrist, C., Thomas, P., Valentin, F., Wilson, D., Wu, C., and Yeats, C. (2007). New developments in the InterPro database. *Nucleic Acids Res.*, 35, D224–228.