BIOINFORMATICS APPLICATIONS NOTE
Vol. 29 no. 14 2013, pages 1811–1812
doi:10.1093/bioinformatics/btt283

Sequence analysis
Advance Access publication May 29, 2013

COPRED: prediction of fold, GO molecular function and functional residues at the domain level

Daniel López and Florencio Pazos*

Systems Biology Department, Computational Systems Biology Group (CNB-CSIC), c/ Darwin, 3. Cantoblanco, 28049 Madrid, Spain

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Only recently the first resources devoted to the functional annotation of proteins at the domain level started to appear. The next step is to develop specific methodologies for predicting function at the domain level based on these resources, and to implement them in web servers to be used by the community. In this work, we present COPRED, a web server for the concomitant prediction of fold, molecular function and functional sites at the domain level, based on a methodology for domain molecular function prediction and a resource of domain functional annotations previously developed and benchmarked.

Availability and implementation: COPRED can be freely accessed at http://csbg.cnb.csic.es/copred. The interface works in all standard web browsers. *WebGL* (natively supported by most browsers) is required for the in-line preview and manipulation of protein 3D structures. The website includes a detailed help section and usage examples.

Contact: pazos@cnb.csic.es

Received on March 11, 2013; revised on April 23, 2013; accepted on May 13, 2013

1 INTRODUCTION

The computational prediction of functional features for newly sequenced proteins is an active field of research because of the pace at which these raw sequences are obtained and the difficulties associated with the experimental functional characterization.

The most widely used strategy for function prediction is to transfer to the query sequence the functional features (e.g. global function and functional residues) available for a homologous protein in a given database or annotation resource (Rentzsch and Orengo, 2009; Valencia, 2005). Because of technical and historical reasons, most of these resources associate functional descriptors to whole proteins, instead of individual domains. Nevertheless, domains are the evolutionary, structural and functional units of proteins and, at least in the case of ‘molecular functions’, these can be individually assigned to particular domains or groups of domains within protein chains. The functional independence of protein domains is exemplified by the fact that many domains found in multi-domain proteins also exist in isolation, as independent proteins, in other organisms (Marcotte *et al.*, 1999). Although associating molecular functions to whole chains without distinguishing the particular domain(s) responsible for them is not an issue for many applications, it can lead to problems in other cases (Lopez and Pazos, 2009). As molecular functions are generally associated with whole chains in the annotation resources, function prediction methods/servers, based on matching against these annotations, predict molecular function at the whole-chain level.

Only recently the first functional annotations at the domain level started to appear (de Lima Morais *et al.*, 2011; Lopez and Pazos, 2009), highlighting the importance of associating molecular functions to domains and the consequences of not doing so in particular cases. Accordingly, prediction methods adapted to these resources and intended for predicting functions at the domain level are required. We have developed one of these methods and showed that its performance in assigning function is higher that a traditional sequence-based search (Lopez and Pazos, 2013). The method is based on a resource of Gene Ontology (Ashburner *et al.*, 2000) molecular function annotations (GO:MF) at the structural domain level (Lopez and Pazos, 2009). In short, profiles are built for the structural domains sharing the same GO:MF term and the same fold (so that they can be structurally aligned). A significant match of a particular region of a query sequence against one of these profiles can be interpreted as a concomitant prediction of fold and GO:MF function for the corresponding domain. Additionally, the conserved positions in these profiles can be interpreted as ‘functional sites’ and consequently transferred to the aligned positions of the query sequence so as to obtain clues on possible functionally important positions. For full details on the methodology and its evaluation see Lopez and Pazos (2013).

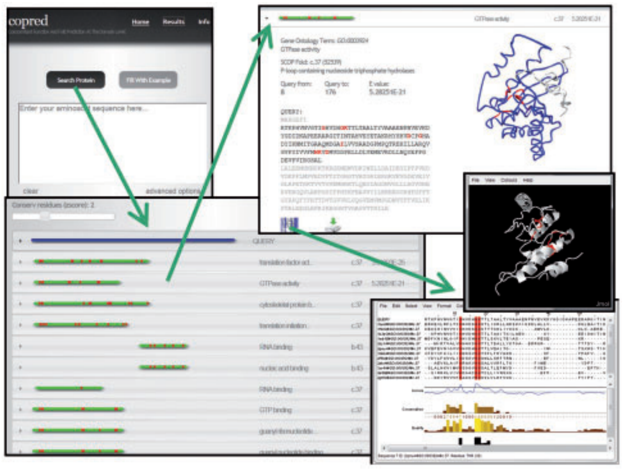
Here, we present COPRED, a web server that allows any user to access this method and generate predictions using, in the simplest case, only the query sequence as input. The predictions can be inspected in a graphical interactive interface and downloaded in a number of standard formats.

2 FUNCTIONALITY AND INTERFACE

Figure 1 shows some representative screenshots of COPRED’s web interface. On the input page, the user can upload the amino acid sequence of the query protein as only input, although an ‘advanced options’ button allows expanding the input form so as to modify some parameters of the method. There is also a ‘fill with example’ button to test the server right away with the example described in the help page.

After a successful run, which normally takes 5–10 s, all the results can be accessed from a single page with different panels that can be expanded/collapsed. The first row represents the query sequence itself (blue in Fig. 1), and expanding it allows

*To whom correspondence should be addressed.



Enter your aminoacid sequence here...

Conserv residues (zscore): 2

translation factor act...	c.37	
GTPase activity	c.37	5.28251E-21
cytoskeletal protein b...	c.37	
translation initiation	c.37	
RNA binding	b.43	
nucleic acid binding	b.43	
RNA binding	c.37	
GTP binding	c.37	
guanyl ribonucleotide...	c.37	

GTPase activity c.37 5.28251E-21

Gene Ontology Term: GO:0003924
GTPase activity

SCOP Fold: c.37 (52539)
P-loop containing nucleoside triphosphate hydrolases

Query from: 8 Query to: 176 E value: 5.28251E-21

QUERY:

Fig. 1. Screenshots of COPRED web interface. Top-left: input page, with links for accessing the help page, inserting an example and for retrieving previous jobs. Bottom-left: unexpanded list of profiles (green) matching different regions of the query sequence (blue). Top-right: expansion of the 'GTPase activity (function)—c.37 (fold)' hit, with details on the region of the query sequence matching that profile, the predicted functional sites, an in-line interactive 3D view of a representative structure of the profile and links for opening a Javlview applet with the multiple sequence alignment associated to the profile, a Jmol applet with the implicit 3D model and the PDB file with that model (bottom-right)

retrieving some data on the input, including a job ID that can be later used to recover the results of previous runs (via the 'Results' option on the input page). The results basically consist of a list of profiles matching different regions of the query sequence, including information on the fold and GO:MF term represented by each of them (Fig. 1). For each hit, a graphical representation of the region of the query sequence matched against the profile is shown, using a color scale (from green to black) to represent the hit's significance (E-value). Consequently, this unexpanded list of hits provides a first overview of the domain composition of the query sequence and the possible folds/functions of its domains. In the example shown in Figure 1, the results are clearly pointing to the presence of two domains: an N-terminal one, associated with the c.37 SCOP fold ('P-loop nucleotide phosphate hydrolases') and GTP-related GO:MF terms, and the middle domain associated with the b.43-fold ('Reductase/isomerase/elongation factor common domain') and RNA-binding–related GO:MF terms. These features are in agreement with the domain composition, structural folds and functions of this example query protein (Elongation factor Tu). Additionally, a set of red dots represent the predicted functional sites transferred from the conserved positions of the corresponding profile (according to a conservation threshold controlled by the scroll-bar at the top of the page).

Expanding a given item of the hit list provides additional information on the corresponding profile match (Fig. 1), such as detailed information on the GO:MF term and SCOP fold, with links to the corresponding databases. The sequence of the query protein is also shown highlighting the region (domain) matching that particular profile and the transferred functional residues. There is also an interactive representation of the 3D structure of a representative member of that profile. This can be manipulated, e.g. zoomed, rotated and so forth. In this 3D view, the region aligned against the query and the conserved residues are highlighted. There is also a link for opening a Jalview (Waterhouse *et al.*, 2009) applet showing the multiple sequence alignment associated with the profile, which also includes the query sequence (Fig. 1). The implicit 3D model for that region (domain) of the query protein is also shown in a Jmol applet (www.jmol.org). Conserved positions are highlighted in both representations. This implicit 3D model, which can be also downloaded as a standard PDB file from another link, contains the backbone atoms of a representative template from the profile with the residues renamed to those of the query sequence.

3 CONCLUSION

Most proteins, especially in eukaryotic organisms comprise multiple domains (Apic *et al.*, 2001). It is important to adapt the function prediction workflows (annotation databases, methods and so forth), mainly developed under a '1-chain-1-function' paradigm, to this reality. Although many existing resources are slowly starting to adopt a 'domain-centric' view, COPRED is the first server specifically devoted to this task. The main advantages of this server are its ease of use and the fact that it provides a first glimpse of the domain structural and functional features of the problem sequence in a concomitant way. The main limitation of the COPRED server at this point is the relatively small scale of the profile database it uses, compared with other resources. Nevertheless, these profiles can be built from any database of functionally annotated structural domains. Therefore, the server can be in principle expanded with larger sets domain annotations.

ACKNOWLEDGEMENTS

The authors thank David S. León (Plant Molecular Genetics Department, CNB) for testing the system and providing feedback and the members of the Computational Systems Biology Group (CNB-CSIC) for comments and suggestions.

Funding: This work was partially funded by the BIO2010-22109 project of the Spanish Ministry of Science and Innovation.

Conflict of Interest: none declared.

REFERENCES

Apic,G. *et al.* (2001) Domain combinations in archaeal, aubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.

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

de Lima Morais,D.A. *et al.* (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.*, **39**, D427–D434.

Lopez,D. and Pazos,F. (2009) Gene ontology functional annotations at the structural domain level. *Proteins*, **76**, 598–607.

Lopez,D. and Pazos,F. (2013) Concomitant prediction of function and fold at the domain level with GO-based profiles. *BMC Bioinformatics*, **14**, S12.

Marcotte,E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Rentzsch,R. and Orengo,C. (2009) Protein function prediction—the power of multiplicity. *Trends Biotech.*, **27**, 210–219.

Valencia,A. (2005) Automatic annotation of protein function. *Curr. Opin. Struct. Biol.*, **15**, 267–274.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.