BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3382–3383
doi:10.1093/bioinformatics/btm521

*Sequence analysis*

# PfamAlyzer: domain-centric homology search

Volker Hollich[1] and Erik L.L. Sonnhammer[1,2,*]

[1]Department of Cell and Molecular Biology, Karolinska Institutet, S-171 77 Stockholm and [2]Stockholm Bioinformatics Center, Stockholms Universitet, S-106 91 Stockholm, Sweden

Received on June 12, 2007; revised on September 22, 2007; accepted on October 13, 2007

Advance Access publication October 31, 2007

Associate Editor: Martin Bishop

**ABSTRACT**

**Summary:** PfamAlyzer is a Java applet that enables exploration of Pfam domain architectures using a user-friendly graphical interface. It can search the UniProt protein database for a domain pattern. Domain patterns similar to the query are presented graphically by PfamAlyzer either in a ranked list or pinned to the tree of life. Such domain-centric homology search can assist identification of distant homologs with shared domain architecture.

**Availability:** PfamAlyzer has been integrated with the Pfam database and can be accessed at http://pfam.cgb.ki.se/pfamalyzer.

**Contact:** Erik.Sonnhammer@ki.se

## 1 INTRODUCTION

Protein domains are regarded as building blocks of proteins as well as evolutionary units. Databases such as Pfam (Finn *et al*., 2006) or SCOP (Andreeva *et al*., 2004) contain domain families based on sequence or structure similarity. With the steady growth of these databases, the need for efficient access methods rises likewise.

PfamAlyzer was created with the aim to enable in-depth analysis of the wealth of information contained within the Pfam database. So far, the various services available on the Pfam homepages have focused on single protein domains or clans of domains instead of complete domain architectures. PfamAlyzer incorporates the majority of the features present at the Pfam database sites and adds tools for the study of domain architectures as well as their appearance throughout evolution.

We here propose domain-centric homology searching for exploring distantly related homologs. It is conducted in two steps: first predicting the Pfam-A domain architecture for a given sequence, and second searching the UniProt database (Bairoch *et al*., 2005) for similar architectures.

## 2 PFAMALYZER APPLET

The PfamAlyzer Java applet has four views that focus on different aspects of the Pfam database. We describe these views in the typical usage order, starting with HMM searching followed by domain query, and finally the domain and SwissPfam browsing.

*To whom correspondence should be addressed.

### 2.1 Protein HMM search

The 'Protein HMM search' view lets the user enter a protein sequence, which is analysed for Pfam-A domains. This is achieved using the HMMER (http://hmmer.janelia.org/) package, which is also applied for generating Pfam full alignments from curated seeds. The user may choose to alter the predefined cutoff values, thus augmenting or shrinking the number of matches. After submitting the query a job ID is immediately displayed. Once the computations are finished, this ID allows accessing the results not only within the current PfamAlyzer session but also at a later point in time.

### 2.2 Domain Query view

PfamAlyzer enables comprehensive study of domain architectures in the 'Domain Query' view. Using a user-friendly graphical query language, arbitrary domains may be freely combined on the screen. PfamAlyzer uses Drag & Drop functionality as an intuitive query formulation approach. Specific domains can be excluded from the search and tolerated gaps between domains may be specified as domains or amino acids. PfamAlyzer allows defining lower and upper bounds for each gap. Thus, the search is not necessarily restricted to exactly defined architectures, but allows some user-defined freedom. Optionally, the search can be limited to specific taxonomic groups, e.g. Chordata. All UniProt proteins with a Pfam domain are queried for the selected architecture. The outcome is presented either in a list fashion or as a species distribution. Here, the proteins are shown as leaves on the species tree. Subtrees may be in-/excluded from visualization, to simplify the analysis of large result sets.

### 2.3 Browse SwissPfam view

The 'Browse SwissPfam' view holds information about whole protein sequences and the domains they contain. Both the manually curated and annotated Pfam-A domains and the uncurated Pfam-B domains derived from the ProDom database are included. Links to the UniProt database provide means for further information gathering.

### 2.4 Browse Pfam view

The 'Browse Pfam' view shows comprehensive information about all Pfam-A families and clans. This includes description, technical parameters for the profile HMM generation, as well as the seed and full alignments. Additionally, the NIFAS tool can

 


**Fig. 1.** The 'Domain Query' view of PfamAlyzer. The upper right frame in the left window shows the query architecture. Additional information for Pfam-A domains is available through context menus as well as links to other resources. The lower right frame presents the result as a list. The manually curated Pfam-A domains are displayed as ovals, whereas automatically curated Pfam-B domains are shown as rectangles. The left frame is used to model the query architecture. Domains can be selected and inserted in the query by Drag & Drop. The right window shows an excerpt from the species distribution. The sequences are ordered on a tree according to species. Subtrees may be in or excluded for in-depth analysis. More help is available through the help menu.

be started for Pfam-A families to explore the phylogenetic tree built from a given domain alignment. Various links to other databases such as SCOP and HOMSTRAD are available as well.

## 3 EXAMPLE

We have created an example that demonstrates the features of PfamAlyzer in a domain-centric homology search. This mini-tutorial starts by searching Pfam with a query sequence. Choose 'Protein HMM search' from the menu buttons and paste in the sample sequence (PLCG1_BOVIN) available at http://pfam.cgb.ki.se/pfamalyzer/sample_sequence. Note that to enable sequence pasting from the clipboard, the applet must be granted the extended access rights. The Protein HMM search performs merged local and glocal searches, but it is possible to run only one type of search to save time (default glocal). Press the submit button to start the domain architecture prediction. When the job is run, the result can be fetched by pressing the 'Retrieve result' button and will be displayed in the lower frame. For the sample sequence, there is support for the domains SH3_1 and SH3_2 at the same positions. For the Domain Query, we will replace the two domain predictions with the more general SH3 clan. Click the right mouse button on the displayed domain architecture to open the context menu and start the Domain Query by selecting 'Find architecture'. Delete the two SH3 domains by dragging them away from the domain architecture composition. Select the SH3 clan from the list on the left (under menu-item clans) and drag it into position between the second SH2 and the PI-PLC-Y domains. Finally, start the domain architecture query with the search button (see result in Figure 1). Changing to the Species distribution tab shows which phyla the found proteins are present in. The query architecture can easily be modified by changing the domains with Drag & Drop. The online help at the PfamAlyzer site describes the software in more detail.

## 4 WEB SERVICE

Besides the user-friendly Java applet, we also offer the Domain Query functionality as a Web Service. The Web Service for the Domain Query is a J2EE application that runs on the JBoss application server. We have created a code example for download at http://pfam.cgb.ki.se/pfamalyzer/example.zip. The example is a project archive for the Eclipse IDE (http://www.eclipse.org/downloads/index.php) to use with the JBoss plugin (http://labs.jboss.com/jbosside). With exception of the test class with the main program, all other classes have been created automatically by JBoss IDE using the description file published at http://www.cgb.ki.se:8080/pfam/WSFacadeService Port?wsdl.

*Conflict of Interest*: none declared.

## REFERENCES

Andreeva,A. *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Finn,R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.