

Protein Ontology and Domain Interaction Database and Web Interface

YoGa Tool



Powered by



Abstract

YoGa Tool associates a database and a web interface. It was designed to allow analysis of interactions between protein ontology stored in GO database and protein domains stored in Prosite. We reached this goal by gathering data from three databases: Ensembl, GO and Prosite. Our web interface provides several ways of search and allows specific and complex queries. We showed that this new way of looking into these relationships uncovers unsuspected connections but also supports former knowledge. We intend to improve our tool to allow broader searches such as multi-species results and automatic updates of data from Ensembl, GO and Prosite.

Introduction

Proteins are architects of life; their role might be the greatest one in life mechanism: any phenotype can be explained by proteins however their function can be described on various scale. You can focus on its global function or break its structure into smaller parts, called protein domains, each one of them is achieving a specific task. There are many database dedicated to these kind of knowledge. Gene Ontology [1] stores molecular function of proteins and their role in biological process; Prosite [2] describes proteins in term of domains and functional sites. Although every piece of information is available, only few studies address the potential interaction between global function and protein domains. We constructed a database of all the transcripts associated with manually annotated proteins, along with their function, and their functional sites. Analysis of this new database will allow us to uncover interactions between the characteristics of these proteins. It will also allow us to further understand how specific domains could lead to specific functions and on the opposite how specific domains are required for specific functions.

Materials and Methods

Transcript and access codes Extraction (via Biomart):

First query was designed to gather transcripts and gene access codes (TAC and GAC) of each transcript including alternative spliced ones. We extracted

them from ensembl48 / NCBI36 (Homo sapiens) [3] using Biomart [4]. This query included two filters: one to keep only protein coding transcripts and another one to eliminate those with automatic ontology annotations (every GO Evidence Code (GOEC) except Inferred from Electronic Annotation (IEA)).

Second query allowed us to match these selected TAC to GO access code (GOAC), GO description (GOD) and GOEC.

Third query was designed to match TAC to Prosite access code (PAC).

Gene Ontology Extraction:

Thanks to the SQL version of GO database available on GO website we were able to match GOAC with the type of ontology (molecular function, biological process or cellular component). Indeed we wanted to focus on molecular function but it is beyond Biomart scope.

Prosite Extraction:

We developed a parser in python [5] to process Prosite flat file in order to integrate it into our SQL database. We only kept PAC and Prosite description (PD) to reduce the size of our database.

Database and Web Interface:

Our database was created in MySQL through PhpMyAdmin and our web interface was developed with Notepad++ using PHP language to access the MySQL database.

Results

Web Interface:

We can divide our web interface into two parts: the search page and results pages. The search page allows database access through several entries. User can perform search on TAC, GAC, GOAC, PAC and even

more complex searches with direct SQL queries (Fig 1). A basic help is also provided. It contains examples of queries, database design (names of fields and tables) and our report.

The screenshot displays the YoGa Tools web interface. At the top center is a logo featuring a white silhouette of a person in a yoga pose next to the text 'YoGa Tools' in a stylized font. Below the logo, the interface is divided into two main sections. On the left, there are four search input fields, each with a 'Search' button: 'Transcript ID' (Ensembl Transcript Accession Number), 'Gene ID' (Ensembl Gene Accession Number), 'Prosites ID' (Prosites Accession Number), and 'GO ID' (Gene Ontology Accession Number). These fields are labeled A, B, C, and D respectively. On the right, there is a large text area for a 'Complex Query', labeled E. Below the text area are two buttons: 'Display' and 'Txt Export'. At the bottom of the interface, there is a footer that says 'Please Help !' and 'Best Viewed in 1280x800. Powered By' followed by logos for 'e! prosites GO' and 'The Gene Ontology'.

Figure 1. Home page of YoGa Tool provides search on a) TAC, b) GAC, c) PAC, d) GOAC and e) complex query.

TAC, GAC, PAC and GOAC search results display information from the database, internal links to our database (by clicking on any access code) and external links towards original database: Ensembl, GO and Prosites (Fig 2). An easy color code is used to display the ontology type: blue for Molecular Function, purple for Cellular Component and green for Biological Process. In GOAC and PAC results, the search term is highlighted in red to improve visibility.

Considering that complex search scope could not be predetermined, this results page was designed dynamically. It means that the display will adjust according to user query. However if users don't change field names of our database, results will be enhanced with internal and external links based on access codes. This kind of query also provides text export in a format readable by classical softwares such as spreadsheet (Excel) and statistical tools (R, SAS).

Transcript : ENST00000225995 ↗	
Gene : ENSG00000108852 ↗	
GO:0004385 : guanylate kinase activity (TAS) - MF ↗ GO:0005515 : protein binding (ISS) - MF ↗ GO:0005624 : membrane fraction (TAS) - CC ↗ GO:0005887 : integral to plasma membrane (TAS) - CC ↗ GO:0009986 : cell surface (IEP) - CC ↗ GO:0007165 : signal transduction (TAS) - BP ↗	PS00856 : Guanylate kinase-like signature ↗

Figure 2. TAC search results displays TAC, GAC, GOAC, GOD, GOEC, PAC and PD

SELECT GOQUERY.*, prosite.proid,pattern.prodesc FROM prosite,pattern,(SELECT go.trid,go.goid,go.godesc FROM go,term WHERE term.goid=go.goid AND term.term_type="molecular_function") AS GOQUERY WHERE pattern.proid=prosite.proid AND prosite.trid=GOQUERY.trid GROUP BY prosite.proid ORDER BY goid,proid Txt Export				
trid	goid	godesc	proid	prodesc
ENST00000264932 ↗	GO:0000104 ↗	succinate dehydrogenase activity	PS00504 ↗	Fumarate reductase / succinate dehydrogenase FAD-binding site
ENST00000315985 ↗	GO:0000158 ↗	protein phosphatase type 2A activity	PS01024 ↗	Protein phosphatase 2A regulatory subunit PR55 signature 1

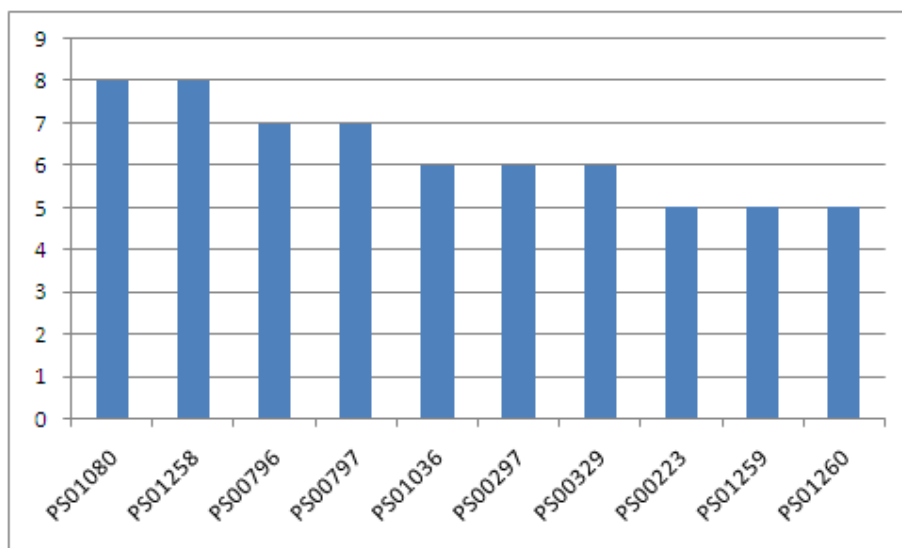
Figure 3. Complex results provides dynamic display and txt export

Statistical Analysis:

Graph 1 shows the number of occurrences of domains PAC associated with a specific GOAC within our database. This graph was created with Excel using our complex query text export.

Table 1 associates PAC from Graph 1 with their description and frequencies. The query used was the following:

```
SELECT
COUNT(prosite.proid),prosite.proid
,pattern.prodesc
FROM (SELECT * FROM go WHERE
goid="GO:0006916")AS ONE, prosite,
pattern
WHERE ONE.trid=prosite.trid AND
prosite.proid=pattern.proid
GROUP BY prosite.proid ORDER BY
COUNT(prosite.proid) DESC
```



Graph 1. Frequencies of Prosite Domains associated with GO anti-apoptosis (GO:0006916) among YoGa Database

Frequencies	PAC	Description
8	PS01080	Apoptosis regulator, Bcl-2 family BH1 motif signature
8	PS01258	Apoptosis regulator, Bcl-2 family BH2 motif signature
7	PS00796	14-3-3 proteins signature 1
7	PS00797	14-3-3 proteins signature 2
6	PS01036	Heat shock hsp70 proteins family signature 3
6	PS00297	Heat shock hsp70 proteins family signature 1
6	PS00329	Heat shock hsp70 proteins family signature 2
5	PS00223	Annexins repeated domain signature
5	PS01259	Apoptosis regulator, Bcl-2 family BH3 motif signature
5	PS01260	Apoptosis regulator, Bcl-2 family BH4 motif signature

Table 1. PAC from Graph 1 and their description

Discussion

Through our web interface we provided a new approach concerning interaction between ontology and protein domains therefore allowing us to uncover relationships between specific active sites and global functions. Our interface allows digging further into these relationships through internal links and complex queries. For example it is very easy to search which protein domains are required to achieve specific functions.

In addition, complex queries associated with text export allowed us to uncover unsuspected relationships between functions and domains. According to Graph 1 and Table 1 we can suspect a strong interaction between anti-apoptosis function and 14-3-3 proteins. Moreover it confirms that Bcl-2 family and heat shock family plays a role in this process. We were also able to characterize huge matter of ATP binding (196 Go occurrences and 218 Prosite occurrences) and Nucleic acid binding (402 Go occurrences and 328 Prosite occurrences) in proteins activity (data not represented here).

Perspectives

Although we designed our database rather strictly, broader analysis could be done by including non-manual annotations, analyzing biological process and cellular interaction or performing multi-species analyses.

Furthermore the tool we created, yet powerful could be improved in many ways. We think that the first upgrade should be an automatic update script designed to daily replicate Ensembl, Go and Prosite. The major difficulty being parsing Prosite flat file considering that Ensembl and Go are already in SQL format. Another improvement could be to provide the possibility to perform typical complex queries easier than through Complex search. We could imagine that searching on a list of transcripts instead of a single one will automatically search for shared characteristics such as GO molecular function or Prosite domains. Finally, a dynamic graphical representation of results could be an interesting challenge and provide an easier way to interpret the results.

This article was written by Yoan L'HOSTIS-JACQUEMIN on a project executed with the collaboration of Gaël GORET

Bibliography

[1] **Consortium, The Gene Ontology.** Gene Ontology: tool for the unification of biology. *Nature Genet.* 2000, Vol. 25.

[2] **Hulo N., Bairoch A., Bulliard V., Cerutti L., Cuche B., De Castro E., Lachaize C., Langendijk-Genevaux P.S., Sigrist C.J.A.** The 20 years of PROSITE. *Nucleic Acids Res.* 2007 Nov 14.

[3] **T. J. P. Hubbard, B. L. Aken, K. Beal1, B. Ballester1, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Ensembl** 2007. *Nucleic Acids Res.* 2007, Vol. 35.

[4] **Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W.** BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 2005, Vol. 21, 16.

[5] <http://python.org/>. *Python Home.* [Online]