BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 8 2006, pages 997–998
doi:10.1093/bioinformatics/btl050

*Genome analysis*

# The DIMA web resource—exploring the protein domain network

Philipp Pagel[1,2,†,*], Matthias Oesterheld[2,†], Volker Stümpflen[2] and Dmitrij Frishman[1,2]

[1]Department of Genome Oriented Bioinformatics, Technical University of Munich, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany and [2]Institute for Bioinformatics/MIPS, GSF, Research Center for Environment and Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany

Received on November 10, 2005; revised on January 16, 2006; accepted on February 7, 2006
Advance Access publication February 15, 2006
Associate Editor: Christos Ouzounis

## ABSTRACT

**Summary:** Conserved domains represent essential building blocks of most known proteins. Owing to their role as modular components carrying out specific functions they form a network based both on functional relations and direct physical interactions. We have previously shown that domain interaction networks provide substantially novel information with respect to networks built on full-length protein chains. In this work we present a comprehensive web resource for exploring the Domain Interaction MAp (DIMA), interactively. The tool aims at integration of multiple data sources and prediction techniques, two of which have been implemented so far: domain phylogenetic profiling and experimentally demonstrated domain contacts from known three-dimensional structures. A powerful yet simple user interface enables the user to compute, visualize, navigate and download domain networks based on specific search criteria.

**Availability:** http://mips.gsf.de/genre/proj/dima

**Contact:** p.pagel@gsf.de

The modular architecture of proteins has been a focus of interest for a long time (Pawson and Nash, 2003). Researchers have made significant efforts to elucidate structure and function of conserved protein domains as building blocks of the proteome.

Today, conserved domains are seen as functional entities which are reused in the context of different proteins, similar to modular components of electronic devices. Some of them represent binding modules while others are associated by functional links.

Based on the well-known method of protein phylogenetic profiling, we recently introduced the idea of domain phylogenetic profiling and demonstrated its utility for linking functionally related and physically interacting proteins (Pagel *et al*., 2004).

Here we present a novel web resource which integrates data sources describing or predicting links among conserved protein domains resulting in a domain interaction map (DIMA). The user is provided with convenient facilities for searching for individual domains, navigation through the network and visualization of subnets. So far, two data sources have been integrated: domain phylogenetic profiling and domain contact evidence from iPFAM (Finn *et al*., 2005). Future releases of the resource will gradually add more data sources and prediction methods.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

*Choosing parameters.* Like most prediction methods, domain phylogenetic profiling depends on a set of parameters which the user can modify in a comprehensive preference form. The most basic parameter is the selection of organisms to be included in the phylogenetic profiles. As of writing, a maximum of 209 completely sequenced public genomes are stored in the PEDANT database (Riley *et al*., 2005) which underlies our profiling technique. The user can choose any number and combination of genomes as input data for the profiling procedure. To ease selection, we offer predefined groups such as 'eukaryota' or 'archaea' which can be selected or deselected with a single mouse click.

The resulting profiles are filtered by information content (Shannon's entropy) according to a user-defined threshold in order to exclude low-information profiles from the analysis. Finally, 'neighboring' profiles are determined based on one of the three available distance/similarity measures: bit distance (Hamming distance), entropy-weighted bit distance and mutual information.

The choice of parameter combinations has a great impact on the resulting predictions. For example, profiling only bacterial proteomes will automatically exclude domains only found in eukaryotes. At the same time, all domains which are present in all used genomes will receive a phylogenetic profile consisting of all '1s' and consequently have zero information content. These will be filtered out if an entropy threshold is used.

The iPFAM data represents bona fide experimental evidence—not predictions—and requires no parameter selection.

*Searching for domains.* The simplest case is the task of finding a specific domain and searching for its immediate neighbors in the network. We currently offer different ways of finding the desired domains. The user can either enter a PFAM (Bateman *et al*., 2004) or InterPro (Mulder *et al*., 2005) accession ID for the domain(s) of interest or conduct a text search using the common domain name or parts of its description.

Finally, the user may be interested in domain relations of some or all of the domains found in a specific protein. In that case, the amino acid sequence of the protein can be used as a query to search the PFAM domain database using the hmmer software (Durbin *et al*., 1998). PFAM domains significantly matching the input sequence will then be used to automatically query the system for domain relations.

The results of these queries are reported in table format including domain IDs, short descriptions and evidence from the individual methods (Fig. 1).

*Computing entire networks.* While searching specific domains is likely to be the most common task needed by users, some



 

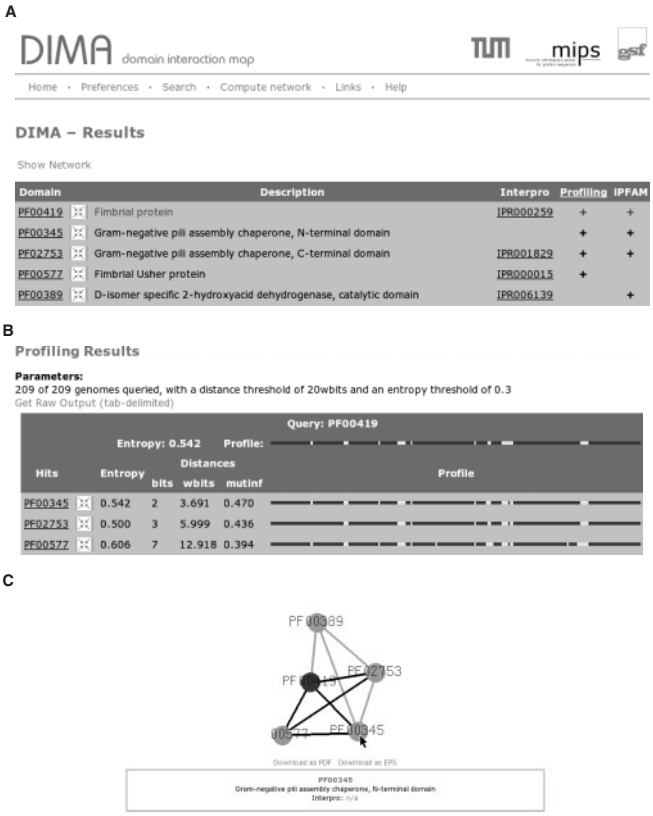


**Fig. 1.** Example for DIMA results. **(a)** Main result table with PFAM and InterPro IDs plus short domain description. The last two columns indicate the methods/data supporting the association. **(b)** Detailed view of domain profiling results. **(c)** Graphical representation of a local domain neighborhood.

researchers may be more interested in global features of the entire domain network generated using a certain set of parameters. For these users we provide the option of having an entire network computed and returned to them by email. Networks are returned as a tab separated table for easy parsing and browsing. The decision to use off-line computation was based on the significantly higher computing times compared with individual queries. Nevertheless, response times are currently very pleasing even for entire networks.

*Results from individual methods.* All results from individual methods (currently only two) can be inspected separately. In the case of domain phylogenetic profiling we provide a graphical representation of the profiles as well as basic parameters like profile entropy and distance. The complete raw output of the profiling tool is also available.

*Visualization.* All generated 'neighborhood' graphs can be viewed graphically (Fig. 1c). We offer two layout variants in order to meet different needs. Force directed layout simulates physical properties of nodes and edges. Nodes repel each other owing to simulated electrical charge, while edges exert attractive spring forces. This algorithm usually gives good results even for large graphs. Hierarchical graph layout is more suitable for small graphs and capitalizes on a more structured layout allowing easier identification of multiple edges and node identification. Graph images can be explored by placing the mouse cursor on a node: Name, ID and a short description of the node will be displayed in a separate info-box. The graphics can be downloaded in three different formats (PNG, EPS and PDF) for off-line use—e.g. in a publication. Graph layout is performed using the powerful program AiSee (http://www.aisee.com).

Clicking a node starts a query of all its neighbors making it easy to navigate the network. The same feature is available in the table representation of the results.

Entire domain networks can grow very large and thus often cannot be handled by the layout program in reasonable time, if at all. At the same time, individual nodes and edges cannot be clearly distinguished in very large graphs—especially if the connectivity is high. Therefore, we currently do not offer visualization of entire networks.

*Performance.* An average query for a single domain using default parameters is answered in less than half a second by the back-end (Pentium-III 800 MHz, 512 Mb RAM). Therefore, in a realistic situation, the delay between hitting the search button and getting results is predominantly determined by the overall load of the web server and the connection speed. In our tests, response times varied between <1 s and up to 5 s.

Computing an entire network even with conservative parameters takes at least 20 s. Once very permissive thresholds for information content and distance are chosen the networks grow significantly larger and hence take longer to build. Therefore, we chose to queue requests for whole networks on the server and deliver the results by email upon completion.

*Conclusion.* The DIMA Web server is the only currently available resource which combines computational predictions of functionally coupled protein domains with experimental data on domain interactions. The quality of the derived domain interaction networks is poised to improve as the number of sequenced genomes and the coverage of the PFAM database grow.

## ACKNOWLEDGEMENTS

We thank Louise Riley and Werner Mewes for careful reading of the manuscript and helpful suggestions. This work was funded by a grant from the German Federal Ministry of Education and Research (BMBF) within the BFAM framework (031U112C).

*Conflict of Interest*: none declared.

## REFERENCES

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Finn,R. *et al.* (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Mulder,N. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.

Pagel,P. *et al.* (2004) A domain interaction map based on phylogenetic profiling. *J. Mol. Biol.*, **344**, 1331–1346.

Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Riley,M. *et al.* (2005) The PEDANT genome database in 2005. *Nucleic Acids Res.*, **33**, D308–D310.