BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 9 2010, pages 1260–1261
doi:10.1093/bioinformatics/btq119

Structural bioinformatics
Advance Access publication March 17, 2010

ArchSchema: a tool for interactive graphing of related Pfam domain architectures

Asif U. Tamuri[1] and Roman A. Laskowski[2,*]

[1]Division of Mathematical Biology, The National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Burkhard Rost

ABSTRACT

Summary: ArchSchema is a Java Web Start application that generates a dynamic 2D network of related Pfam domain architectures. Each node corresponds to a different architecture (shown as a sequence of coloured boxes) and indicates whether any 3D structural information is available in the PDB. Satellite nodes can show either the UniProt codes or the PDB codes of proteins having the given architecture. Search options allow search by UniProt code or Pfam domain identifier, and results can be filtered by domain, organism, or by selecting only proteins in the PDB.

Availability: ArchSchema can be freely accessed at http://www.ebi.ac.uk/Tools/archschema

Contact: roman@ebi.ac.uk

Received on January 11, 2010; revised on March 8, 2010; accepted on March 11, 2010

1 INTRODUCTION

Domains are the fundamental evolutionary building blocks of proteins (Bork, 1992), some occurring particularly frequently, allowing the assembly of myriad proteins from a modest stock of units (Bashton and Chothia, 2002, 2007). A given sequence of domains defines a protein's domain architecture. Many different proteins, in terms of sequence and/or function, can have the same, or very similar, architectures.

The Pfam database (Bateman *et al.*, 2002) identifies protein domains on the basis of sequence alignments, hidden Markov models and manual curation. It can list all the architectures containing any domain and, in turn, all sequences belonging to any of these architectures. There are several other resources that exploit the Pfam data, focusing on domain architectures. PfamAlyzer (Hollich and Sonnhammer, 2007) provides a graphical user interface that allows searches for domain patterns, listing the architectures that match. CDART (Geer *et al.*, 2002) and DAhunter (Lee and Lee, 2008) both identify homologous proteins based on the similarity of their domain architectures. The web server d-Omix (Wichadakul *et al.*, 2009) can show static 2D graphs of similarities in domain architectures for a user-submitted set of protein sequences.

Here, we describe ArchSchema that shows related Pfam domain architectures as a dynamic graph rather than as a list of matches. It provides a clear 2D visualization of how the architectures are related. A further advantage over other methods is that it flags any nodes which have associated 3D structural information in the Protein Data Bank, PDB (Berman *et al.*, 2003).

2 METHODS

ArchSchema is a Java Web Start application that uses data derived from the Pfam, UniProt (The UniProt Consortium, 2010) and PDBsum (Laskowski, 2009) databases. It makes use of two freely available Java libraries, namely the Prefuse visualization toolkit for generating the dynamic graphs and the BrowserLauncher by Eric Alberts of Stanford University. Similarities between architectures are computed using a simple Needleman and Wunsch alignment (Needleman and Wunsch, 1970), and nearest-neighbours are joined by edges, as described in the program's online documentation.

3 RESULTS

ArchSchema can be launched from the ArchSchema home page (http://www.ebi.ac.uk/Tools/archschema) or from links within PDBsum. It can also be downloaded and installed locally for use when needed.

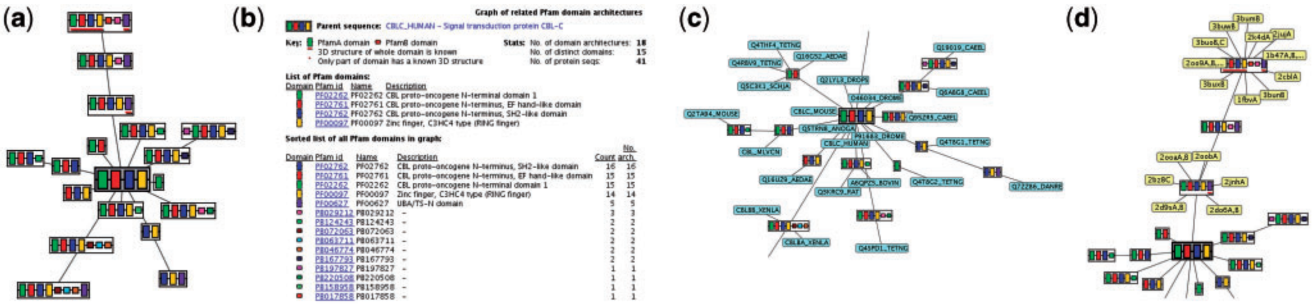
3.1 Example: human CBL-C protein

Figure 1a shows an initial ArchSchema graph for human CBL-C protein (UniProt code CBLC_HUMAN), a signal transduction protein, which has no structures in the PDB. The protein consists of four Pfam domains and its domain architecture is indicated by the larger node with a grey background at the centre of the graph. This is termed the 'parent architecture'. The graph contains 18 related architectures in all, each having one or more of the parent's four domains; 11 of the architectures contain all four domains. Details are provided in a 'data panel' (Figure 1b) that lists all the domains on the graph in descending order of their frequency.

One can pan and zoom around the graph using click-and-drag operations and can 'switch on' satellite nodes to show either the UniProt sequences associated with each node (Figure 1c), or the PDB codes of structures belonging to each node (Figure 1d). Where there is a huge number of additional nodes, only 50 are shown and are coloured pink to indicate a particularly popular architecture. The lengths of the edges joining the nodes can be adjusted using a slider bar, while another slider bar allows you to prune the outer nodes in very large networks.

Clicking on a node of interest in the graph lists its UniProt sequences in the data panel. These are hyperlinked to the UniProt database and, any that have 3D structures in the PDB, are hyperlinked to PDBsum. This provides a means of tracking down protein structural information from related domain architectures.

*To whom correspondence should be addressed.



Graph of related Pfam domain architectures

🟩🟥🟦🟨 **Parent sequence:** CBLC_HUMAN – Signal transduction protein CBL-C

Key: 🟩 PfamA domain ● PfamB domain
3D structure of whole domain is known
Only part of domain has a known 3D structure

Stats:	
No. of domain architectures:	**18**
No. of distinct domains:	**15**
No. of protein seqs:	**41**

List of Pfam domains:

Domain	Pfam id	Name	Description
🟩	PF02262	PF02262	CBL proto-oncogene N-terminal domain 1
🟥	PF02761	PF02761	CBL proto-oncogene N-terminus, EF hand-like domain
🟦	PF02762	PF02762	CBL proto-oncogene N-terminus, SH2-like domain
🟨	PF00097	PF00097	Zinc finger, C3HC4 type (RING finger)

Sorted list of all Pfam domains in graph:

Domain	Pfam id	Name	Description	Count	No. arch.
🟦	PF02762	PF02762	CBL proto-oncogene N-terminus, SH2-like domain	16	16
🟥	PF02761	PF02761	CBL proto-oncogene N-terminus, EF hand-like domain	15	15
🟩	PF02262	PF02262	CBL proto-oncogene N-terminal domain 1	15	15
🟨	PF00097	PF00097	Zinc finger, C3HC4 type (RING finger)	14	14
🟪	PF00627	PF00627	UBA/TS-N domain	5	5
●	PB029212	PB029212	-	3	3
●	PB124243	PB124243	-	2	2
●	PB072063	PB072063	-	2	2
●	PB063711	PB063711	-	2	2
●	PB046774	PB046774	-	2	2
●	PB167793	PB167793	-	2	2
●	PB197827	PB197827	-	1	1
●	PB220508	PB220508	-	1	1
●	PB158958	PB158958	-	1	1
●	PB017858	PB017858	-	1	1

Fig. 1. Examples of ArchSchema graphs for the signal transduction protein CBLC_HUMAN. (**a**) Initial graph showing all domain architectures sharing one or more of the four domains belonging to CBLC_HUMAN (the 'parent architecture', the slightly enlarged node with a grey background in the middle of the graph). The coloured boxes correspond to different Pfam domains: tall boxes for Pfam-A domains and short boxes for Pfam-B domains. Architectures with proteins that have structural models in the PDB are annotated by a red line under the domains that have been structurally determined: a long red line indicates that the structure encompasses at least 90% of the domain's residues, a short red line that the structure covers <90% of the domain, and no line means there is no (or very little, i.e. <50%) structural information available. A continuous line joining two or more domains means that one or more 3D structures contain all the underlined domains together, whereas if the line between the domains is broken it means that the domains are found only in separate PDB entries. (**b**) The key to the graph from the 'data panel' listing the Pfam domains in the parent architecture plus all domains on the plot in descending order of occurrence. (**c**) A small part of the network, centred on the parent node, with satellite nodes added (coloured light blue) showing the UniProt sequences having each architecture. (**d**) The top part of the graph in (a), with the satellite nodes (light yellow) representing PDB entries for proteins having the given domain architecture. Note that the relative positions of the nodes adjust whenever any satellite nodes are added or removed as the network's layout is dynamically optimized.

For example, the red line in topmost node in Figure 1a indicates that one or more 3D structures of the first four domains exist—precisely the four domains of our CBLC_HUMAN protein.

3.1.1 Refining the search A search panel, in a separate tab, allows you to refine your search or to start a new search—perhaps to further explore any of the architectures on the current graph. If the parent sequence contains several different domains, you can search for architectures containing any one, or more, of these domains. You can filter the search by restricting it to a specific organism. And finally, you can limit the search to just those proteins for which there is structural information in the PDB.

3.1.2 Excessive data In some cases, the initial search may return a vast number of related architectures. For example, the human tumour suppressor protein BRCA1 returns 2379 architectures that contain at least one of its two Pfam domains, corresponding to 10 361 different UniProt protein sequences. This would be far too much data to easily make sense of if displayed at once. Thus, ArchSchema uses two strategies to focus on the most interesting part of the network whenever the number of architectures exceeds 150. First, the Pfam-B domains in all the architectures are replaced by a single dummy domain, code PB000000. In the case of BRCA1_HUMAN, this 'collapses' the number of unique architectures to 954. These are further filtered according to their computed similarity to the parent architecture. Only the closest 150 architectures, or thereabouts, are retained and shown on the plot.

ACKNOWLEDGEMENTS

We would like to thank Rob Finn and Marija Buljan for useful comments on the program.

Conflict of Interest: none declared.

REFERENCES

Bashton,M. and Chothia,C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.*, **315**, 927–939.

Bashton,M. and Chothia,C. (2007) The generation of new protein functions by the combination of domains. *Structure*, **15**, 85–99.

Bateman,A. *et al*. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Berman,H. *et al*. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.

Bork,P. (1992) Mobile modules and motifs. *Curr. Opin. Struct. Biol.*, **2**, 413–421.

Geer,L.Y. *et al*. (2002) CDART: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.

Hollich,V. and Sonnhammer,E.L.L. (2007) PfamAlyzer: domain-centric homology search. *Bioinformatics*, **23**, 3382–3383.

Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.

Lee,B. and Lee,D. (2008) DAhunter: a web-based server that identifies homologous proteins by comparing domain architecture. *Nucleic Acids Res.*, **36**, W60–W64.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Wichadakul,D. *et al*. (2009) d-Omix: a mixer of generic protein domain analysis tools. *Nucleic Acids Res.*, **37**, W417–W421.

The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.