Bioinformatics, 32(15), 2016, 2384–2385
doi: 10.1093/bioinformatics/btw140
Advance Access Publication Date: 12 March 2016
Applications Note

Sequence analysis

TreeDom: a graphical web tool for analysing domain architecture evolution

Christian Haider[1,2], Marina Kavic[1,2] and Erik L.L. Sonnhammer[1,*]

[1]Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Solna SE-17121, Sweden and [2]FH OÖ – University of Applied Sciences Upper Austria, Hagenberg 4232, Austria

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 18, 2015; revised on February 12, 2016; accepted on March 3, 2016

Abstract

Summary: We present TreeDom, a web tool for graphically analysing the evolutionary history of domains in multi-domain proteins. Individual domains on the same protein chain may have distinct evolutionary histories, which is important to grasp in order to understand protein function. For instance, it may be important to know whether a domain was duplicated recently or long ago, to know the origin of inserted domains, or to know the pattern of domain loss within a protein family. TreeDom uses the Pfam database as the source of domain annotations, and displays these on a sequence tree. An advantage of TreeDom is that the user can limit the analysis to N sequences that are most similar to a query, or provide a list of sequence IDs to include. Using the Pfam alignment of the selected sequences, a tree is built and displayed together with the domain architecture of each sequence.

Availablility and implementation: http://TreeDom.sbc.su.se

Contact: Erik.Sonnhammer@scilifelab.se

1 Introduction

Multi-domain proteins have evolved by insertions or deletions of distinct protein domains. Tracing the history of a certain domain combination can be important for functional annotation of multi-domain proteins, and for understanding the function of individual domains (Forslund and Sonnhammer, 2012). In order to analyze the evolutionary history of the domains in multi-domain proteins it is desirable to inspect the sequence tree for each domain together with the domain architecture of the sequences.

A few graphical tools exist that can combine trees with domain architectures. NIFAS (Storm and Sonnhammer, 2001) is a Java applet based on outdated Java libraries that does not work in many current browsers. Although it is a general-purpose tool, it has mainly been used as a plugin to the Pfam database (Finn *et al.*, 2014), but this is no longer supported. Another discontinued tool is the TreeDomViewer (Alako *et al.*, 2006). DoMosaics (Moore *et al.*, 2014) is a Java applet package that includes a panel with a tree and unproportional domain architectures. However, it requires substantial manual work to produce such a view for a given case, which limits its usability. The PhylomeDB database (Huerta-Cepas *et al.*, 2014) for complete catalogs of gene phylogenies has a similar graphic display that also contains features other than domains. It is however limited to the phylogenies in PhylomeDB and not for general purpose. The same is true for the InParanoid database (Sonnhammer and Östlund, 2015) that shows domain architectures and tree for proteins in an orthologue group. Finally, the TreeFam database (Schreiber *et al.*, 2014) of animal phylogenetic trees has a tree and domain architecture graphic which is built on a reusable BioJS.Tree (Yachdav *et al.*, 2015) JavaScript component.

2 Implementation

We created a website called TreeDom that lets the user choose a protein family and sequences from Pfam, then fetches them, computes a tree and displays it together with the Pfam domain architectures. It uses the BioJS.Tree component, version 1.0.0, for the graphical display. The back-end is implemented as PHP scripts to enter the query and a C++ program to fetch the alignment and domain architectures from Pfam's Web services, and to compute the tree. The tree is built using the Neighbour-Joining (NJ) method with uncorrected distances and is rooted at the midpoint, i.e. where the sum of the branch lengths of the two subtrees branching out from the root are as close as possible.

Because protein families are very large nowadays, often with thousands of members, it is not practical to show all of them. The TreeDom website thus offers several ways to extract the sequences of interest only. It is based on the Pfam database and can be run in two modes:

Single query:

- Input: 1 family ID; 1 member query sequence ID, N
- Build NJ tree for N closest sequences using Pfam's alignment

Multiple query:

- Input: 1 family ID; N member sequence IDs
- Build NJ tree for all provided sequences using Pfam's alignment

TreeDom is close in functionality to NIFAS. The main advantages of TreeDom are that it works in any web browser, and that it can limit the analysis to N sequences using two different methods. Such input limitation was never implemented for NIFAS which therefore could not handle families with more than ~500 members. A major drawback with NIFAS was that the applet only showed one simultaneous tree, but in order to analyze the evolution of different domains it is essential to compare their trees visually. This drawback is removed by TreeDom by running it in multiple web browser windows.

TreeDom was optimized for speed by a C++ implementation of the back-end which completes within a few seconds for most families when asking for less than 50 sequences. The two time consuming steps are fetching (i) the alignment and (ii) the domain architectures from the Pfam server. The former can be sped up by choosing a representative proteome dataset (Chen *et al.*, 2011). These redundancy-reduced datasets also promote sequence diversity in the tree and are usually preferable over the full alignments.

3 Example

An example of domain architecture evolution analysis of NUDIX hydrolases is shown in Figure 1. The aim is to look for domain transfer, domain loss and domain duplication. Starting from NUD13_HUMAN, most of the closest homologs have the same domain architecture [NUDIX-like]-[zf-NADH-PPase]-[NUDIX] but some exceptions exist where the NUDIX-like or zf-NADH-PPase domains are missing or truncated. Only one of these, H2SB95_TAKRU, is annotated as a fragment, suggesting that the NUDIX domain is more essential than the other two domains. Mammals and oviparous animals are fairly well separated clades in these trees. The coelacanth which lies in between is unusual because has two proteins (*_LATCH). The NUDIX-like domains of these proteins have diverged about as much as between species, suggesting an old duplication (Fig. 1b), whereas the NUDIX domains (Fig. 1a) are almost identical. Possible explanations for this are gene conversion, or that the gene duplication is recent but the NUDIX-like domain is not under selection. Among the shown species there is no evidence for domain transfer or domain duplication. In fact, 18 of the 20 closest homologs to the NUDIX and NUDIX-like domains are the same, suggesting that this domain architecture is very stable in vertebrates.

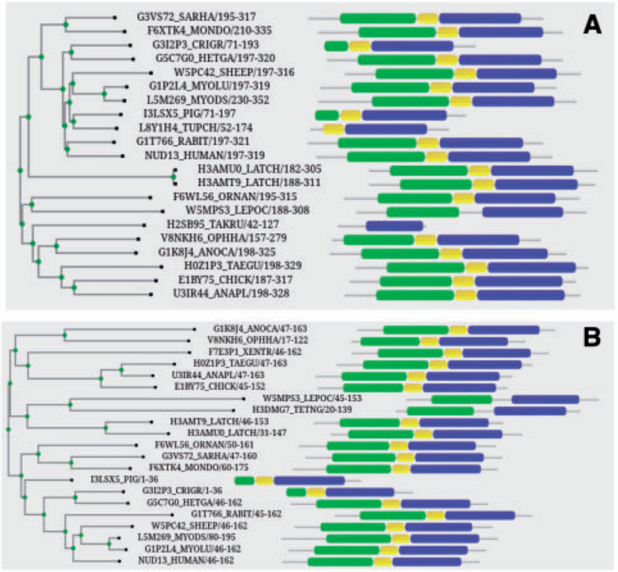

Fig. 1. (**A**) TreeDom output using the query NUDIX domain (PF00293), NUD13_HUMAN as single query, 20 closest sequences and RP55 (representative proteomes at 55% co-membership). The domains are: blue/right:NUDIX, green/left:NUDIX-like (PF09296), yellow/middle: zf-NADH-PPase (PF09297). (**B**) Same as A but instead using the NUDIX-like domain for building the tree. On the website, the sequences and domains are hyperlinked to the corresponding Pfam entries (Color version of this figure is available at *Bioinformatics* online.)

Acknowledgement

We thank Fabian Schreiber for help with the BioJS.Tree software.

Funding

This work has been supported by the Erasmus programme.

Conflict of Interest: none declared.

References

Alako,B.T. *et al.* (2006) TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure. *Nucleic Acids Res.*, **34**, W104–W109.

Chen,C. *et al.* (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**, e18910.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Forslund,K. and Sonnhammer,E.L. (2012) Evolution of protein domain architectures. *Methods Mol. Biol.*, **856**, 187–216.

Huerta-Cepas,J. *et al.* (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.

Moore,A.D. *et al.* (2014) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics*, **30**, 282–283.

Schreiber,F. *et al.* (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.*, **42**, D922–D925.

Sonnhammer,E.L. and Östlund,G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239. 2015,

Storm,C.E. and Sonnhammer,E.L. (2001) NIFAS: visual analysis of domain evolution in proteins. *Bioinformatics*, **17**, 343–348.

Yachdav,G. *et al.* (2015) Anatomy of BioJS, an open source community for the life sciences. *Elife 2015*, **4**, e07009.