BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 2 2014, pages 282–283
doi:10.1093/bioinformatics/btt640

Sequence analysis

Advance Access publication November 12, 2013

DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins

Andrew D. Moore[1,*,†], Andreas Held[†], Nicolas Terrapon[1,†], January Weiner, 3rd[2] and Erich Bornberg-Bauer[1,*]

[1]Institute for Evolution and Biodiversity, Hüfferstrasse 1, Westphalian Wilhelms-University Münster, 48147 Münster, Germany, and [2]Max Planck Institute for Infection Biology, Chariteplatz 1, 10117 Berlin, Germany

Associate Editor: Burkhard Rost

ABSTRACT

Summary: DoMosaics is an application that unifies protein domain annotation, domain arrangement analysis and visualization in a single tool. It simplifies the analysis of protein families by consolidating disjunct procedures based on often inconvenient command-line applications and complex analysis tools. It provides a simple user interface with access to domain annotation services such as InterProScan or a local HMMER installation, and can be used to compare, analyze and visualize the evolution of domain architectures.

Availability and implementation: DoMosaics is licensed under the Apache License, Version 2.0, and binaries can be freely obtained from www.domosaics.net.

Contact: radmoore@uni-muenster.de or e.bornberg@uni-muenster.de

Received on January 7, 2013; revised on September 13, 2013; accepted on October 26, 2013

1 INTRODUCTION

Protein evolution can be treated as a series of domain-wise changes that alter their domain arrangement (or architecture), the sequence of a proteins' constituent domains. The majority of novel arrangements can be explained by rearranging domains and processes such as fusion, fission and terminal domain loss are thought to be the primary forces at play (Moore *et al*., 2013). The signatures of domains are stored in specialized domain databases such as Pfam (Punta *et al*., 2012), SUPERFAMILY (de Lima Morais *et al*., 2011) or InterPro (Hunter *et al*., 2012). Existing tools that exploit domain content for evolutionary analysis focus on specific aspects, primarily the identification of homologous proteins, and are often limited to a single domain definition database. Services are either provided in the form of web servers such as d-Omix (Wichadakul *et al*., 2009) or DAhunter (now WDAC) (Lee and Lee, 2008), or as standalone applications such as PfamAlyzer (Hollich and Sonnhammer, 2007), Nifas (Storm and Sonnhammer, 2001) or XDOM (Gouzy *et al*., 1997). The recently described tool ArchSchema (Tamuri and Laskowski, 2010) takes a domain-graph approach to illustrate related domain arrangements. However, none of these tools combine annotation, analysis, search and visualization into a single framework.

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

DoMosaics collates all steps from domain annotation, homology search, analysis of architectural evolution to creation of customizable publication-ready figures into a single tool.

2 METHODS

DoMosaics is a JAVA program and requires JAVA 1.6 or better. Phylogenetic trees can be constructed based on sequences remotely using SOAP services or locally using distances between domain arrangements in the form of domain edit or Jaccard distances. Tree computation makes use of the Phylogenetic Analysis Library by Drummond and Strimmer (2001); tree visualization loosely uses procedures described by Griebel *et al*. (2008). Searches for similar domain arrangement by domain string alignments are performed using the RADS/RAMPAGE web service (Terrapon *et al*., 2013). Finally, DoMosaics allows for content-dependent annotation using CODD (Terrapon *et al*., 2009).

3 RESULTS

The program can be started directly from the DoMosaics web page using JAVA Web Start. Alternatively, a precompiled JAVA archive file (JAR) can be downloaded. An example dataset and the program documentation are included in the program.

3.1 Example: *Drosophila* aminoacyl-tRNA synthetases

We applied DoMosaics to a simple analysis of the *Drosophila* aminoacyl-tRNA synthetases (see Fig. 1). We extracted FlyBase defined orthologs of the *Drosophila melanogaster* protein FBpp0083898 and loaded the corresponding fasta file into a new DoMosaics project. Domain annotation was conducted in two steps. First, all sequences were annotated using the remote hmmpfam scan provided by the InterProScan service for which DoMosaics provides a graphical user interface. Second, domain annotation was derived based on a local installation of hmmscan against Pfam-defined models; in this second scan we used a high *E*-value cutoff of 2.0 as opposed to the model-defined gathering thresholds used by InterProScan. Domain annotation reveals four unique arrangements (see Fig. 1a). All arrangements contain the C-terminal domains ProRS-C_1 and HGTP_anticodon. Three arrangements contain N-terminal deletions; the arrangement found in *Drosophila sechellia* lacks the WHEP-TRS repeat array and the C-terminal region of the tRNA-synt_2b domains. Traces of the latter are only found by loosening *E*-value; however, such traces are also recognizable in the domain dotplot (see Fig 1b). The context-dependent annotation using CODD and

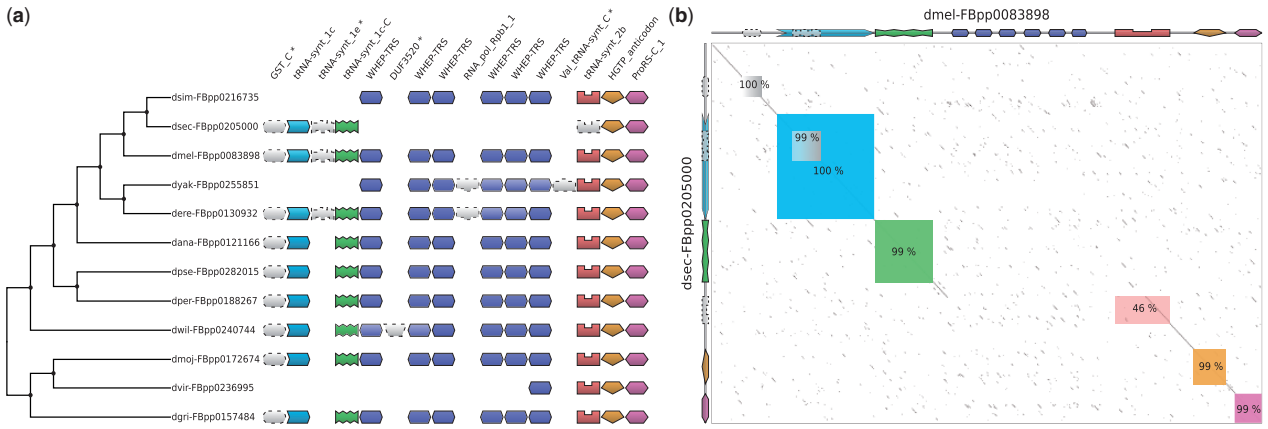

Fig. 1. Drosophila orthologs of aminoacyl-tRNA synthetases. (**a**) Domains were annotated and arrangements were associated with a species tree. Domains were manually aligned; same domains were aligned by comparison of domain sequences. Domains are illustrated as unproportional boxes; dotted white boxes signify CODD domains that match below the model defined threshold, yet can be asserted based on domain co-occurrence (Terrapon *et al*., 2009). Annotation against Pfam models shows four unique arrangements; context-dependent annotation conducted using CODD and loosened *E*-value thresholds show traces of a C-terminal glutathione S-transferase domain in nine arrangements and traces of tRNA-synt_1e, RNA_pol_rpb1_1 and Val_tRNA-synt_C in three, two and one arrangement, respectively. (**b**) Domain dotplot (a dotplot that includes pairwise domain sequence comparisons indicated as percent identity) of the aminoacyl-tRNA synthetases found in *D.melanogaster* and *D.sechellia*. The domain dotplot indicates a deletion of the WHEP-TRS repeat array, possibly also effecting the tRNA-synt_2b domain, of which only traces can be found in *D.sechellia* (figure exported as.svg followed by minor edits: border/color change of putative domains; above-tree domain legend)

relaxed *E*-value thresholds reveals traces of a number of previously unseen domains, such as the C-terminal occurring GST_C domain or the tRNA-synt_1e domain.

Although it remains unclear whether differences correspond to evolutionary events or annotation artifacts stemming from profile bias or failure in gene predictions, such analysis is quickly conducted and allows visual inspection of possible events. All steps, from the annotation to the analysis and visualization can be conducted within minutes.

ACKNOWLEDGEMENT

The authors would like to thank Stefan Rensing, Adam Sardar, Andreas Schüler and three anonymous reviewers for providing valuable feedback.

Funding: This work was supported by the DFG (Deutsche Forschungs Gemeinschaft) grant (BO 2455/4-1 to E.B.B).

Conflict of Interest: none declared.

REFERENCES

De Lima Morais,D.A. *et al.* (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.*, **39**, D427–D434.

Drummond,A. and Strimmer,K. (2001) PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, **17**, 662–663.

Gouzy,J. *et al.* (1997) XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. *Comput. Appl. Biosci.*, **13**, 601–608.

Griebel,T. *et al.* (2008) EPoS: a modular software framework for phylogenetic analysis. *Bioinformatics*, **24**, 2399–2400.

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

Hunter,S. *et al.* (2012) Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

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.

Moore,A.D. *et al.* (2013) Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochim. Biophys. Acta*, **1834**, 898–907.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

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

Tamuri,A.U. and Laskowski,R.A. (2010) ArchSchema: a tool for interactive graphing of related Pfam domain architectures. *Bioinformatics*, **26**, 1260–1261.

Terrapon,N. *et al.* (2009) Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics*, **25**, 3077–3083.

Terrapon,N. *et al.* (2013) Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics* [Epub ahead of print, doi:10.1093/bioinformatics/btt379, July 4, 2013].

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