Protein family analysis at the domain-level

Nicolas Terrapon, Andrew D. Moore and Erich Bornberg-Bauer
Institute for Evolution and Biodiversity, University of Münster

ebb@uni-muenster.de

Abstract: The analysis of protein domains has gained considerable attention over the last years. Many new insights on protein modular evolution, combined with improved domain detection, have paved the way for an integrated analysis of protein families from a domain-centric perspective. We recently released DoMosaics, a JAVA application that facilitates the interactive analysis of protein domain arrangements. DoMosaics combines guided domain annotation, a highly-customisable visualization of arrangements, and a number of analysis tools. It also integrates domain-centric algorithms such as CODD, which is used for the detection of divergent domain occurences that have escaped Pfam thresholds, as well as RADS/RAMPAGE which provides means to search for proteins with a domain arrangement similar to a given query. RADS provides an alignment of domain strings as opposed to amino-acid sequences, while RAMPAGE produces an amino-acid alignment guided by RADS results. Hence, RADS/RAMPAGE produces fast and yet accurate alignments, and associated ranking, of proteins with similar domain arrangements. Together, these tools greatly simplify the domain-centric analysis of protein function, structure and evolution.

1 Introduction

The evolution of gene-encoding proteins is not only driven by mutation, insertion or deletion of single nucleotides, but also involves the rearrangement of larger genomic regions which requently correspond to protein domains. Domains are the smallest structural, functional, and evolutionary units of proteins. They usually vary in length between 100-250 amino-acids, except for short repeat motifs [CGVT03]. Domains became a cornerstone of protein annotation thanks to Hidden Markov Models (HMMs), a powerful approach that captures family diversity, and databases of domain families such as Pfam that cover a large part of the protein universe [FBC$^+$14]. Most proteins contain only one domain, while multidomain proteins represent less than 33% of proteins encoded by genomes [WKCA11]. However, the sequential order of the domains in a protein, or a protein's "domain arrangement", can be subject to recombination and has been shown to be a major factor of evolution and novelty in complex multicellular organisms. In the past 15 years, many studies have provided insights into the underlying mechanisms of modular protein evolution. An important foundation was the observation that many domains are ancient and shared between all organisms, as well as some domain combination with a strong conservation of N- to C-terminus order [AGT01]. In a recent study, we observed that the majority of novel arrangements can be explained by simple rearranging processes such as fusion, fission and terminal domain loss [MGS+13]. These properties notably gave rise to domain-based algorithms such as the Co-occurrent Domain Detection (CODD) which allows detection of divergent domains [TGMB09] based on patterns of domain co-occurrence, or the recent RADS/RAMPAGE approach which can identify and align similar proteins based on their domain content [TWG+14].

2 RADS/RAMPAGE [TWG+14]

A key task in the analysis of protein families is the identification of a protein set which share similar domain arrangements. RADS (Rapid Alignment of Domain Strings) determines the similarity between two proteins by aligning their domain arrangements, using a classical dynamic programming algorithm, and hence does not require any amino-acid sequence information. A key advantage to this approach is the reduction in time complexity: while proteins in UniProt contain on average 324 amino acids, they harbour an average of only 1.5 domains (2.6 for multi-domain arrangements). The second method, RAMPAGE (Rapid Alignment Method of Proteins based on domain ArranGEments) complements RADS and addresses the need for increased sensitivity. RAMPAGE creates global alignments of amino-acid sequences using the domain-wise alignments provided by RADS as a guideline (see Figure 1). RAMPAGE alleviates the problem of aligning single-domain arrangments with RADS and performs with a sensitivity similar to, but significantly faster than, BLAST. We demonstrated that these methods yield biologically meaningful results, which work at a speed that is significantly faster than classical local alignment tools. We provided a fast C-based command-line application for running custom domain-string comparisons, a web interface for querying UniProt with Pfam and a command-line JAVA application for querying the web interface in batch mode, which can also be used as a JAVA library for programmatic access. To satisfy the need for a tool which simplifies the analysis of protein families by uniting these new approaches and by offering powerful visualization abilities, we developed DoMosaics [MHT+14].

3 DoMosaics [MHT+14]

DoMosaics is a Java application that unifies protein domain annotation, domain arrangement analysis and visualization in a single tool. It simplifies the analysis of protein families by unifying disjunct procedures based on partly inconvenient command-line based applications and complex analysis tools. DoMosaics provides easy, GUI-based access to domain-annotation services such as InterPro, and can work without internet connection after downloading HMMER binaries, at http://hmmer.janelia.org, and HMM libraries from public domain databases as Pfam [FBC+14]. It can be used to analyze the change of domain arrangements along a phylogenetic tree, construct domain-guided dotplots and domain co-occurence graphs, and perform divergent domain detection with CODD [TGMB09], and retrieve proteins with similar domain arrangements using queries to Uniprot with

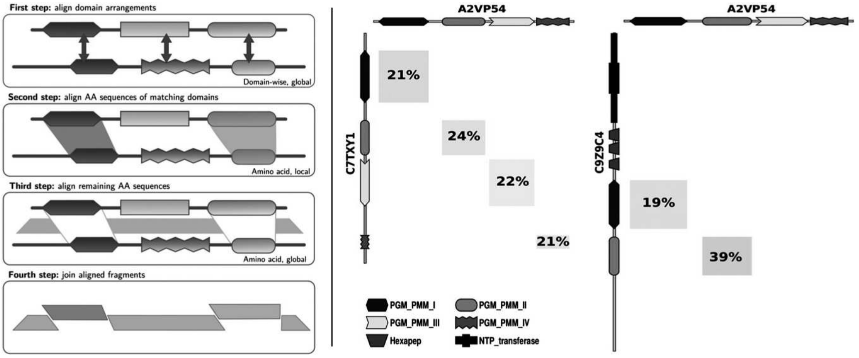
First step: align domain arrangements
Domain-wise, global
Second step: align AA sequences of matching domains
Amino acid, local
Third step: align remaining AA sequences
Amino acid, global
Fourth step: join aligned fragments
A2VP54
C7TXY1
21%
24%
22%
21%
A2VP54
C9Z9C4
19%
39%
PGM_PMM_I
PGM_PMM_II
PGM_PMM_III
PGM_PMM_IV
Hexapep
NTP_transferase

Figure 1: Principle of the domain-driven amino-acid alignment by RAMPAGE (left panel). The right panel illustrates a query A2VP54 with RADS that recognizes the similar arrangement of C7TXY1, while BLAST fails to find it with an e-value threshold of 0.1 but finds the distant arrangment of C9Z9C4 at 10^{-6}.

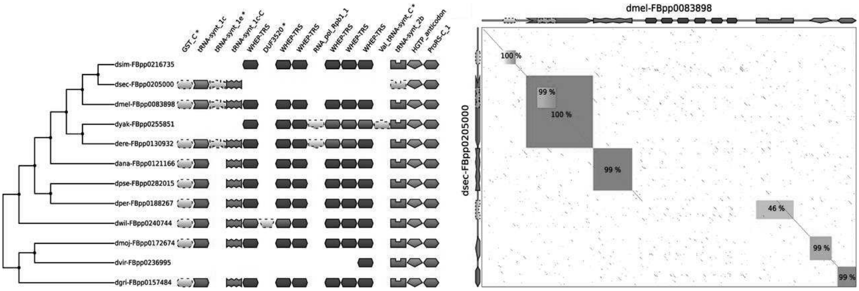

Figure 2: DoMosaics view of one protein family and domain-wise dotplot.

RADS/RAMPAGE [TWG+14]. Finally, DoMosaics allows for highly-customisable visualization, and can export high-quality, publication-ready images of protein domain arrangements.

References

[AGT01] G. Apic, J. Gough, and S.A Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of molecular biology*, 310(2):311–325, 2001.

[CGVT03] C. Chothia, J. Gough, C. Vogel, and S.A Teichmann. Evolution of the protein repertoire. *Science*, 300(5626):1701–1703, 2003.

[FBC+14] R.D Finn, A. Bateman, J. Clements, et al. Pfam: the protein families database. *Nucleic acids research*, 42(D1):D222–D230, 2014.

[MGS+13] A.D. Moore, S. Grath, A. Schüler, A.K. Huylmans, and E. Bornberg-Bauer. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochimica et Biophysica Acta-Proteins and Proteomics*, 1834(5):898–907, 2013.

[MHT$^+$14] A.D. Moore, A. Held, N. Terrapon, J. Weiner, and E. Bornberg-Bauer. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics*, 30(2):282–283, 2014.

[TGMB09] N. Terrapon, O. Gascuel, É. Maréchal, and L. Bréhélin. Detection of new protein domains using co-occurrence: application to Plasmodium falciparum. *Bioinformatics*, 25(23):3077–3083, 2009.

[TWG$^+$14] N. Terrapon, J. Weiner, S. Grath, A.D. Moore, and E. Bornberg-Bauer. Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics*, 30(2):274–281, 2014.

[WKCA11] M. Wang, C.G. Kurland, and G. Caetano-Anollés. Reductive evolution of proteomes and protein structures. *Proceedings of the National Academy of Sciences*, 108(29):11954–11958, 2011.