Footprints of modular evolution in a dense taxonomic clade

Andrew D. Moore & Erich Bornberg-Bauer
Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, University of Muenster

radmoore@uni-muenster.de, ebb@uni-muenster.de

Abstract: True novelty, of any form, is rare. Most systems, including a number of biological systems, can be reduced to a set of of core units which are reused in varying contexts. These core units can be seen as modules, and their harboring system as modular. Here, we explore various aspects of modularity in protein evolution within a dense clade of 20 arthropods. By employing a simple model of protein evolution, we study how the rearrangements of domains - the modules of protein evolution, structure and function - creates novelty in few steps and at surprising speeds. We find that we can explain between 64% - 81% of all novel protein domain arrangements, and that arrangements that cannot be explained contain curious patterns of domain repeats. Furthermore, we explore the speed of module turnover - the frequency of domain gain and loss - and find that while only few new domains occur, they spread swiftly and seem associated with environmental adaptation.

1 Introduction

A primary factor in the evolution of proteins is the rearranging of protein domains, their functional, structural and evolutionary modules. Using modular rearrangement, functional diversification can occur without the formation of novel domains, simply by adding, removing or rearranging domains in proteins [MBE+08]. Previous studies have illustrated that, in particular along the metazoan lineage, increased rates of domain rearrangement can be found [MBE+08]. Indeed, while the number of identified domains grows very slowly, the number of combinations of these domains continues to grow with no end in sight [Lev09].

As opposed to the often slow variation at the sequence level, events such as gene fusion/fission or the shuffling of exons, which are among the genetic protagonists driving modular domain recombination [BFB10], can swiftly produce selectable phenotypes [RH12, PGWL10]. While a series

of mutations can govern selectable phenotypes, a number of mutations remain unseen to the eye of selection. In contrast, large events such as the fusion of two genes is likely to produce a phenotype, some of which may even be favored by selection [RH12]. Autonomously functioning domains used in a modular system, where functionalities can be recombined easily, provide a powerful mechanism for evolutionary innovation.
From numerous previous studies we know that the dominant mechanisms creating novel arrangements are gene duplication, fusion and terminal losses [WBBB06]; that age, function and structure of a domain do not influence their versatility [WMBB08] and that strings of domains are well suited for designing algorithms for homology search [TGW$^+$12].

While rare, evidence for novelty does exist e.g. in the large number of orphan genes, many of which are presumed to be vital for species-specific quirks [KHF$^+$09]. Beyond genes, changes in domain content between species, and species groups, can be observed [ZG11]. This indicates that novel domains do emerge - albeit at low frequency. It seems plausible that certain molecular innovation, such as required in the wake of strong environmental shifts, may be out of reach by the rearrangement of existing domains alone and may require the emergence of novel domains.

We have recently explored various aspects of modular evolution using a small, well described clade of 20 arthropods. In this data set, we have derived branch-specific rates of events in modular evolution and have assessed the evolutionary dynamics and functional impact of changes at the level of the domain repertoire. Beyond the exploration of various aspects of protein evolution, our approach illustrates the strength of domain-based analysis: the great accuracy of HMMs in identifying homologous sequences and the low rates of domain turnover helps capture functional shifts and evolutionary dynamics at a rather coarse grained level and across evolutionary long time scales of tens to hundreds of million years.

2 Results

Within the arthropods, a total of 30 domain are found to be emergent (that is, occur only within this clade) [MBB12]. By functionally annotating all proteins which harbor an emerging domain (1,291 proteins across 20 arthropods), we assessed the functional impact of novel domains. Domains that emerge within arthropods are found significantly more often in terms related to environmental adaptation (e.g. response to heat, drought, UV and other abiotic stresses), than expected by chance (see figure 1).

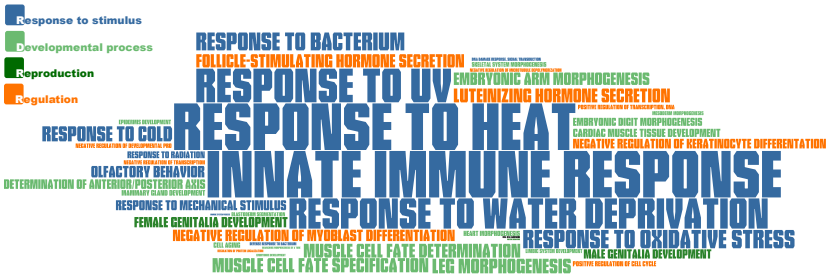

Figure 1: **TermLogo of functional groups with emerging domains.** Over-representation analysis of Gene Ontology terms from proteins which contain at least one emerging domain. The size of the font corresponds to the strength of obtained significance.

The majority of arrangements are unique to one, or very few species (Figure 2). The majority of arrangements are unique to one, or very few species, facilitating a roughly bimodal distribution of shared arrangements. This indicates that modular rearrangement is frequent enough to create a large diversity of arrangements, even in evolutionarily small timescales. Furthermore, while the largest proportion (∼80%) of arrangements shared by all species are single domain proteins, species-specific arrangements tend to be multi-domain indicating that older arrangements tend to be single-domain, while newly formed arrangements are more likely multi-domain.

After ancestral reconstruction of arrangement presence/absence states, we derive rates of arrangement gain for all branches. We then, for each new arrangement, investigate how new arrangements can be formed by recombining ancestral arrangements (e.g a new arrangement (A,B,C) can be formed by the fusion of the ancestral arrangements (A,B) and (C)). We consider the fusion of two arrangements, the fission of an arrangement, as well as the gain or loss of parts of arrangements. We find that we can explain up to 81% of all new arrangements by a single-step event while some new arrangements have conflicting solutions; a total of 64% of all new arrangements have only possible solution.
The evolutionary dynamics of the events are intriguing: while fusion and gain dominate early in the tree, fission and loss frequencies increase over time. A possible interpretation concerns arrangement length: recombination events that give rise to novel (viable) arrangements are likely to act between domains as to not disrupt functional domains. The smaller the number of domains that are present in an arrangement, the lower the

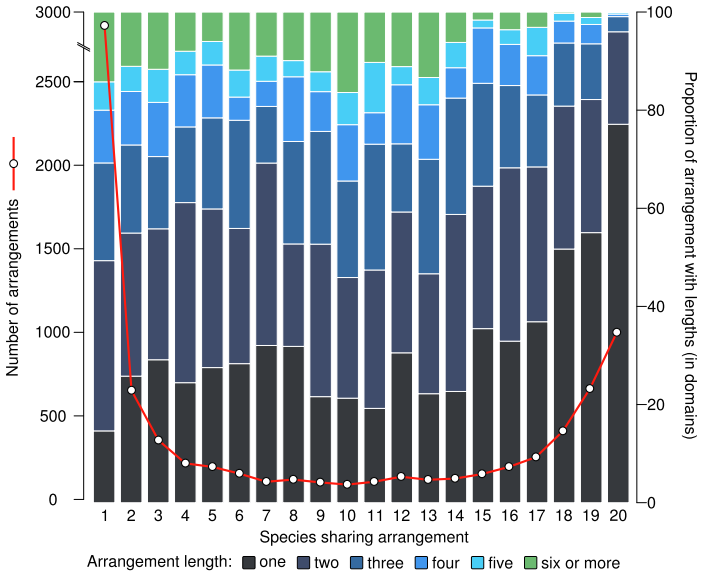

Figure 2: **Unique arrangements and arrangement length in 20 pancrustacean species.** Unique arrangements were grouped by the number of species in which they can be found. The x-axis indicates the number of species which share arrangements, the y-axis indicates the number of arrangements. For each group of shared arrangements, the arrangement length measured as the number of domains was determined and normalized to 100% (z-axis). The red line plot illustrates that the distribution of unique arrangements is roughly bimodal, with the majority of arrangements shared by either few or all species.

chance for successful fission or loss. In contrast, fusion and gain events seem more likely detrimental the longer an arrangement gets.

New arrangements that cannot be explained by one of the considered events contain complex, multi-domain repeat patterns ("supra-repeats") and are significantly enriched in domain-repeats. Such domain-repeats are essential to protein-protein interaction and DNA-binding making them key players in regulatory networks. Beyond the analysis of arthropods, we find that the overall signals in plant species are similar [KBMG12].

In summary, our results provide a detailed account of the mechanisms with which domain rearrangement events create novel proteins, and provide an excellent starting point for further analysis ranging from mathematical modeling to additional cross-species comparisons.

References

[BFB10] Marija Buljan, Adam Frankish, and Alex Bateman. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol*, 11(7):R74, 2010.

[KBMG12] Anna R Kersting, Erich Bornberg Bauer, Andrew D Moore, and Sonja Grath. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol Evol*, Feb 2012.

[KHF+09] Konstantin Khalturin, Georg Hemmrich, Sebastian Fraune, René Augustin, and Thomas C G Bosch. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*, 25(9):404–413, Sep 2009.

[Lev09] Michael Levitt. Nature of the protein universe. *Proc Natl Acad Sci U S A*, 106(27):11079–11084, Jul 2009.

[MBB12] Andrew D Moore and Erich Bornberg-Bauer. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol*, 29(2):787–796, Feb 2012.

[MBE+08] Andrew D. Moore, Åsa K. Björklund, Diana Ekman, Erich Bornberg-Bauer, and Arne Elofsson. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*, 33(9):444–451, Sep 2008.

[PGWL10] Sergio G Peisajovich, Joan E Garbarino, Ping Wei, and Wendell A Lim. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science*, 328(5976):368–372, Apr 2010.

[RH12] Rebekah L Rogers and Daniel L Hartl. Chimeric Genes as a Source of Rapid Evolution in Drosophila melanogaster. *Mol Biol Evol*, 29(2):517–529, Feb 2012.

[TGW+12] Nicolas Terrapon, Sonja Grath, January Weiner, Andrew Moore, and Erich Bornberg-Bauer. Fast Homology Search Using Domain-Architecture Alignment. *JOBIM Conference proceedings*, 2012.

[WBBB06] January Weiner, Francois Beaussart, and Erich Bornberg-Bauer. Domain deletions and substitutions in the modular protein evolution. *FEBS J*, 273(9):2037–2047, May 2006.

[WMBB08] January Weiner, Andrew D Moore, and Erich Bornberg-Bauer. Just how versatile are domains? *BMC Evol Biol*, 8:285, 2008.

[ZG11] Christian M Zmasek and Adam Godzik. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol*, 12(1):R4, Jan 2011.