Functional innovation from changes in protein domains and their combinations

Jonathan G. Lees[1]*+, Natalie L. Dawson[1], Ian Sillitoe[1] and Christine A. Orengo[1]

[1] Institute of Structural and Molecular Biology, Division of Biosciences, University College London, Gower Street, London, WC1E 6BT, UK

*To whom correspondence should be addressed. Tel: 02076793890; Fax: 02076797193; Email: jonathan.lees@ucl.ac.uk

Abstract

Domains are the functional building blocks of proteins. In this work we discuss how domains can contribute to the evolution of new functions. Domains themselves can evolve through various mechanisms, altering their intrinsic function. Domains can also facilitate functional innovations by combining with other domains to make novel proteins. We discuss the mechanisms by which domain and domain combinations support functional innovations. We highlight interesting examples where changes in domain combination promote changes at the domain level.

Introduction

Globular protein domains are structurally compact, independently folding units and can be grouped into sometimes very large homologous superfamilies (Sillitoe et al. 2015). Domain superfamilies can be defined using purely sequence (as in Pfam (Finn et al. 2015)), or sequence combined with structural data (such as used in the CATH and SCOP resources (Sillitoe et al. 2015; Murzin et al. 1995)). As the amount of protein structure data has grown, it has become clear that some homologues can diverge significantly in their structures and functions. In some superfamilies the domain fold can change to such an extent, that it appears similar to domain folds in other superfamilies. Some recent domain structure classifications (e.g. SCOP2 (Andreeva et al. 2014) highlight these structural overlaps, which suggest continuity in some regions of structure space (Edwards & Deane 2015).

Domain sequences of structurally uncharacterised proteins can be assigned to SCOP and CATH structural superfamilies using profile Hidden Markov Models (Eddy 2011) built from sequence alignments of known members (Lam et al. 2015; Oates et al. 2015).Currently CATH and SCOP combined identify ~3,000 superfamilies, which comprise more than 50 million domains and account for nearly 70% of domains in completed genomes. In order to explore functional divergence, superfamilies in CATH have been further sub-classified into ~100,000 more functionally coherent families (FunFams) (Das et al. 2014; Das et al. 2015). Functional sub-classification is performed using a new sequence based clustering algorithm which clusters relatives that share similar sequence preferences (Das et al. 2014). In particular, relatives are sub-classified according to their specificity determining residues, which typically account for specific ligand or protein binding properties. The functional coherence of FunFams has been independently endorsed by the international CAFA assessments (Radivojac et al. 2013; Jiang et al. 2016) and also *in-silico* by examining the accuracy of functional site prediction in FunFams (Das et al. 2014).

Globular domains do not make up all of the functional units of a protein, and in particular, the increasing importance of disordered regions and their functions in proteins has been appreciated (Wright & Dyson 2014). Some disordered regions have been classified as 'constrained disorder' (Bellay et al. 2011) and some Pfam domains are classified as conserved disordered domains (e.g. PF16597) with ~14% of all Pfam domains predicted to be mostly disordered (Tompa & Fersht 2009).

Individual domains can expand their functional repertoire in a number of ways, including residue mutations and loop extensions ((Das et al. 2015) for recent review). Detailed analyses of domain families have suggested that functional properties are largely conserved across the family (Todd et al. 2001) and that these changes mostly promote interactions with novel substrates or new protein partners on different pathways and processes (Das et al. 2015). Additionally, domains are extensively duplicated and combined to produce novel proteins with new **M**ulti-**D**omain **A**rchitectures (MDAs) ((Bashton & Chothia 2007) and (Bornberg-Bauer & Albà 2013) for recent review). The MDA of a protein is the ordered arrangement of its domains (i.e. ABC not equal to ACB) and has been likened to a sentence made up of individual domain-words. Higher level evolutionary units (likened to complex syntax elements in (Scaiewicz & Levitt 2015)) called supra-domains have also been described, consisting of two or three-domains, that are conserved in different proteins and have specific functional and spatial relationships (Vogel et al. 2004) .

Domain data highlights how nature recombines what is there already (Jacob 1977), a recurring theme in many areas of biology (Wagner et al. 2007). It is also clear that some domains perform more valuable functions, or have greater plasticity to adapt to new molecular environments and have been reused more often than others. The most highly duplicated domain superfamilies in CATH i.e. the largest 200 superfamilies, account for two thirds of all domains in CATH and occur in very many different multi-domain contexts (**Figure 1 A-D**). There is a strong positive correlation between the number of domain partners a superfamily has and its number of FunFams (**Figure 1 E**).

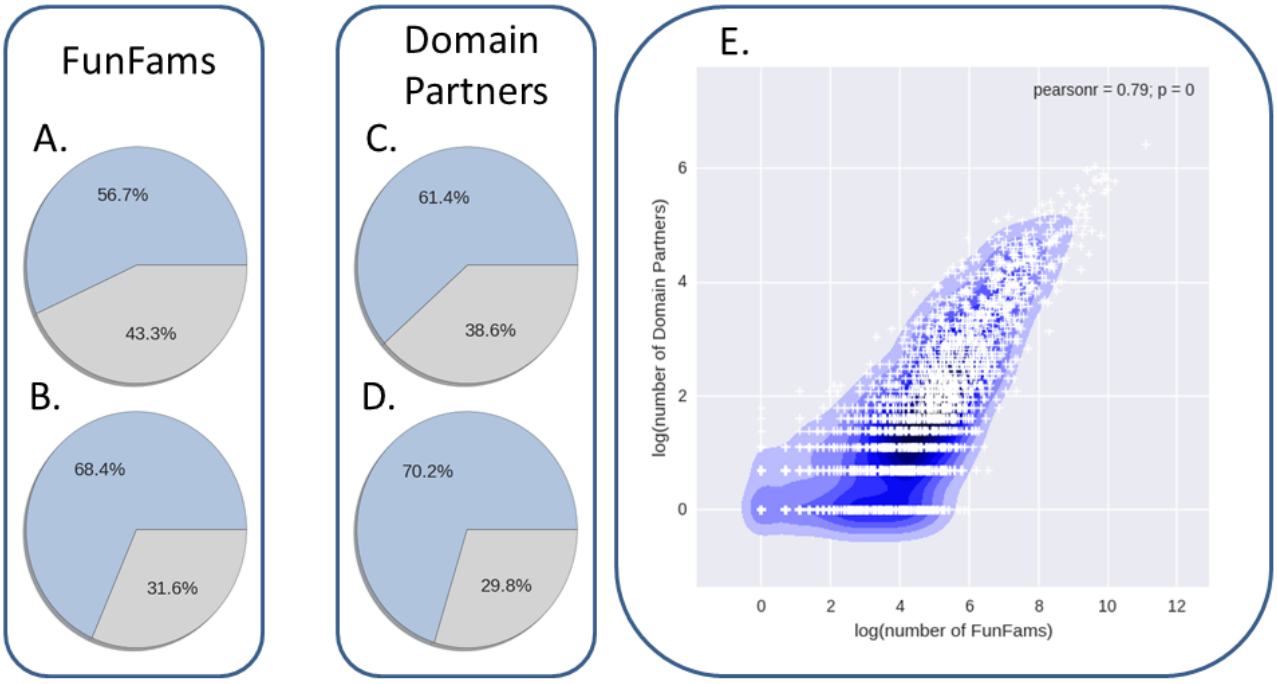

Figure 1*: The first two columns show percentage covered by the largest 100 CATH superfamilies (as measured by the number of FunFams) in blue. In the first column is the percentage of all FunFams covered in A) all pan-compara genomes and B) Metazoan genomes. The second column shows percentages of the number of distinct domain partners C) for all pan-compara genomes and D) just Metazoan genomes. E) There is a strong correlation between the size of a superfamily (as measured by the number of FunFams it has) and the number of domain partners it has.*

A significant proportion of domain superfamilies tend to be widely distributed and using stringent criteria we find that over one third of the CATH domain superfamilies found in human (526 of 1355) are found in all major branches of cellular life (Figure 2). These universal

superfamilies cover an even larger proportion of domain assigned residues in human (Figure 2). Multi-domain combinations show an opposite trend to domain families and tend to be much less consistently widespread (Figure 2). In fact, many lineages show specific MDAs that may help with lineage specific adaptations in a wide variety of contexts from regeneration (Abdullayev et al. 2013) to immunity and symbiosis (Hamada et al. 2013). In the context of evolvability, it has recently been proposed that a subset of MDAs exist within organisms that are frequently built upon to make new MDAs to facilitate organismal adaptions (Hsu et al. 2016). Furthermore, it is known that while some domains are more 'promiscuous' than others (Cohen-Gihon et al. 2011; Bornberg-Bauer & Albà 2013), some gain promiscuity in specific lineages whilst others show a steady increase in promiscuity over long evolutionary periods (Cohen-Gihon et al. 2011).

Undoubtedly there are many clear examples where MDA innovation has been of great importance, such as in building key signalling pathways (Anon 2010). However, it is worth noting a recent study which emphasised the importance of reuse of existing domain architectures to adaptive processes(Sardar et al. 2014).

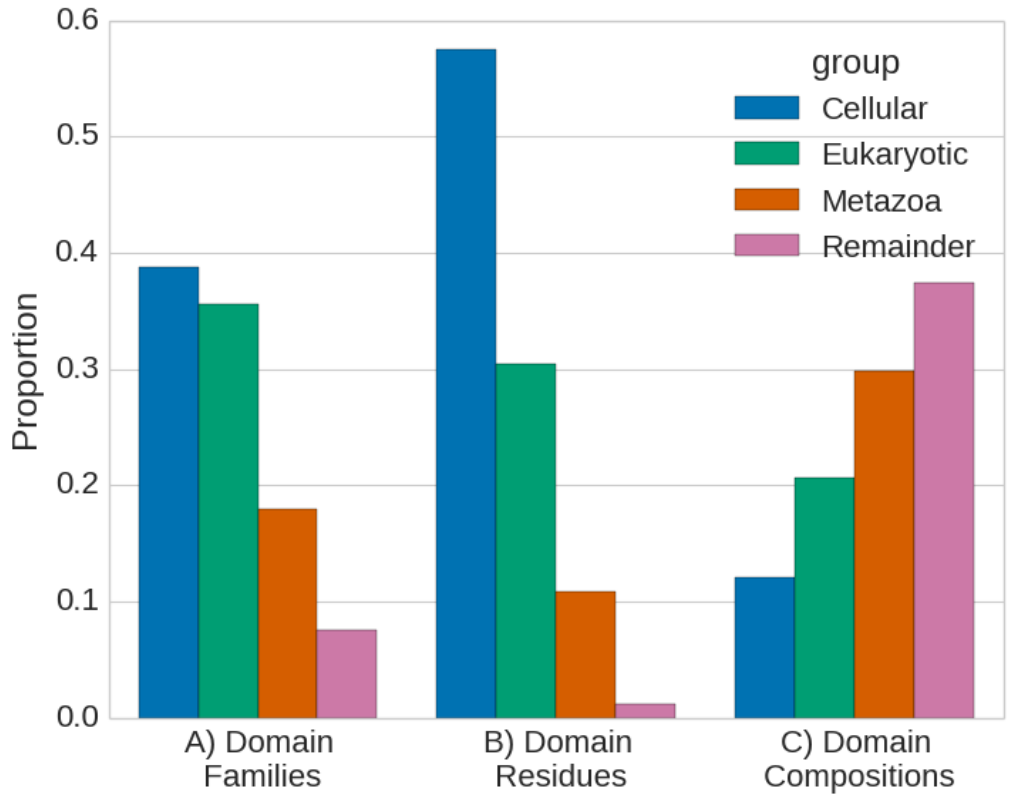

Figure 2. *A) 'Domain Families' shows the proportion of Human domain families found to be widely present at different taxonomic levels. A domain family is assigned to a taxonomic level if it is found to be present in more than 20% of species from each of its main child taxon's (using the NCBI tree) (e.g. for the 'Cellular' group this requires that the domain family be present in >20% from each group of Archaeal, Bacterial and Eukaryotic species). If a domain family is found to occur widely in a more ancient group it is not assigned to a younger group. B) 'Domain Residues' is the proportion of human protein residues assigned to domain families from a given taxonomic level. C) 'Domain Compositions' shows the multi-domain compositions (the set of domains with no respect of order i.e. ABC==ACB) assigned to different taxonomic levels using the same criteria of occurrence rate etc. as for domain families. Note this is purely a representation of how widely occurring different domain combinations are at different taxonomic levels. In the above plots shifting the cut-off from 20% to 50% resulted in broadly similar trends.*

Intra Domain innovations

Some domain superfamilies have diversified hugely during evolution to produce a large array of structures and functions. The 100 most populated families in CATH account for 54% of all domains and comprise 18% and 48% of the different CATH structural groupings (relatives with similar structures i.e. superposing with <5Å RMSD) or functional families (FunFams) respectively. However, the core scaffold (or fold) remains conserved within a superfamily (Sillitoe et al. 2015) and this is supported from what is known from studies on evolutionary site variation (Echave et al. 2016). Certain domains like TIM Barrels support an extremely large number of functions and may have arisen from simple duplications of alpha-beta secondary structures, potentially helping support the initial transition from ribozymes to protein-based enzymes (Goldman et al. 2016). Domains can evolve new functions in a number of ways including changes in individual residues that alter catalytic activities, embellishments to the core fold of a domain that change ligand binding properties and altered interaction partners ((Das et al. 2015) for recent review).

Recent studies of diverse enzyme superfamilies in CATH found changes in more than 50% of catalytic residues in the actives sites of the 70 most diverse enzyme superfamilies (Furnham et al. 2015). Usually a catalytic core is conserved but residues around this can change to enable the chemistry to be performed on a different substrate or to optimise the efficiencies of the enzymes (Furnham et al. 2015; Brown & Babbitt 2014).

Studies of changes in the interfaces used in protein interactions in relatives from diverse CATH superfamilies also showed that relatives could have very different interaction partners and that the interfaces involved were associated with very different parts of the domain surface in different relatives (accounting altogether for most of the surface) (Dessailly et al. 2013).

In many enzyme families the catalytic residues lie within loops that are separated from the core structural scaffold. Mutations in these loops are less likely to destabilise the protein thus allowing changes that can lead to novel specificities or chemistries. Recently, the term 'polarity' has been developed to capture the modularity of the structural scaffold regions and the function mediating residues i.e. the separation of catalytic residues in loops detached from the main scaffold (Dellus-Gur et al. 2013). High polarity (modularity) is found for a number of domains, particularly those with TIM barrel or Rossmann folds and is likely to result in greater evolvability (Dellus-Gur et al. 2013; Toth-Petroczy & Tawfik 2014). The number of functional families in domains with a TIM barrel fold is indeed high (858 different FunFams amongst the pan-compara genomes)

However mutations affecting activities are not always found in the active site, and allostery is an important mechanism affecting binding or activity, as illustrated in a study on Rubisco (Studer et al. 2014). A high throughput mutagenesis study on the photoactive yellow protein PAS domain, which catalogued an array of functional alterations for a complete Ala mutation series showed that many of the mutations affecting function were not present in the active site (Philip et al. 2010). A recent study showed how functional sites (particularly catalytic ones), impose extensive, long range evolutionary constrains over the rest of the protein (Jack et al. 2016).

Despite the innovations observed in highly diverse domain superfamilies, detailed studies of enzyme families have shown that these changes rarely alter the chemistry that the relatives perform (Todd et al. 2001; Furnham et al. 2015; Brown & Babbitt 2014). It appears difficult to engineer new chemistries and changes in the domain are most frequently associated with altered substrate specificities. Furthermore, the catalytic residues performing the chemistry for a given protein are

usually located in a single domain and it's rare for active sites to be formed between domains (Furnham et al. 2012), supporting the 'domain grammar of function' concept. However, interestingly, despite this fact, 87% of chemistries performed by proteins have emerged within existing superfamilies having different ancestral chemistries (Furnham et al. 2015).

Furthermore, there are striking examples of how simple amino acid substitutions can lead to completely different domain functions. A recent example of this is shown for the GK Protein interaction domain (GK_{PID}) which evolved via duplication of a Guanylate Kinase (GK) enzyme before the last common ancestor of Filozoa (Anderson et al. 2016). In Metazoans, the GK_{PID} domain helps orient spindles relative to other cells through its interaction with the Pins protein. Experiments showed only a single amino acid substitution was sufficient to repurpose the GMP binding surface on GK, simultaneously removing enzymatic activity and allowing Pins binding (possibly by changes in conformational occupancy).

By contrast, there are also interesting examples of domains undergoing drastic structural changes whilst maintaining their basic biological functions, with a recent work showing this for a spectrum of extreme structural changes (compared to the typical structural divergences through extending loops as a baseline) including modification (addition /deletion) of secondary structures to the core fold, changes in fold topology and combined spatial and topological transmogrifications(Zhang et al. 2014). This process (diverged structure / conserved function) appeared to be particularly common in 'arms race' like evolutionary scenarios, with for example a change in number of strands in certain kinases affecting antibiotic resistance.

Recent work has shown a surprisingly large amount of domain family loss, with strong functional biases for different lineages (Zmasek & Godzik 2011) and most Eukaryotic lineages showing a net loss of domains. At the emergence of animals there is considerable loss of domains involved in metabolism and gain of domains in regulation with compensation of metabolic losses coming from symbionts (i.e. gut microbes). Punctuated bursts of innovation followed by steady loss is a trend seen in many areas of evolution (Wolf & Koonin 2013).

Novel functions through altered domain combinations

Domains often carry out discrete functions (Finn et al. 2015) and domains from a multi domain protein frequently have the same general function (although with modification or specialisation) as its single domain homologs (Bashton & Chothia 2007). This and other observations has led to the idea of a 'domain grammar' (Bashton & Chothia 2007). It was also noted in this work that sometimes domains can completely alter their function, and this was related to a change in a words meaning. In order to maintain specific functions orthologues generally maintain very similar MDAs (Forslund et al. 2011). Changes in the domain content of a protein can alter its functioning in a number of ways including its enzymatic functions, interaction partners and localisation (Bornberg-Bauer & Albà 2013). Furthermore, when domains are combined into the same gene, it ensures the domains are co-localised and co-expressed.

There are many mechanisms (Marsh & Teichmann 2010) by which a protein can gain domains, with gene fusion appearing to be of greatest importance in Metazoa (Buljan et al. 2010). It is worth noting that several papers, including recently (Triant & Pearson 2015), have discussed how various errors, especially in gene models, can produce incorrect MDA predictions. Proteins can also gain domains with no apparent homologues, so called 'orphan' domains (Bornberg-Bauer & Albà 2013), which could have emerged de-novo from a previously non-coding sequence (Bornberg-Bauer et al. 2015; Bornberg-Bauer & Albà 2013). Methods for identifying orphan domains are helping to better characterise their prevalence and evolutionary history (Bitard-Feildel et al. 2015).

Domains are predominantly gained and lost at termini (Weiner et al. 2006) reflecting both the processes responsible for the losses and the presumed lower chance of structural disruptions (Buljan & Bateman 2009). Domain loss is through processes such as gene fission (Weiner et al. 2006) and recent work has suggested that reversion of gene fusions by such processes is more common than previously thought (Leonard & Richards 2012).

It has also been shown that it is possible for a domain to lose core structural regions leading to a partial but still functional domain in a process termed 'domain atrophy'(Prakash & Bateman 2015). One example, is a TIM barrel domain, comprising less than half the original fold (exposing the large hydrophobic core) which dimerises to bury the hydrophobic surface (Prakash & Bateman 2015).

It is also worth noting that a single protein can dynamically switch its domain content through disorder to ordered transitions (Tompa & Fersht 2009) and in a few cases between different ordered folds (Bryan & Orban 2010). Furthermore, a single gene can give rise to multiple protein isoforms with altered MDAs through processes such as alternative-splicing (Light et al. 2013), potentially causing large shifts in interaction partners (Yang et al. 2016). However, the proportion of protein isoforms that are found in the proteome is unknown and many isoforms remain undetected in proteomics experiments, although homologous exon substitution events (which are likely to be important for modifying an individual domains function (Abascal, Tress, et al. 2015)) are found to be relatively overrepresented at the proteomics level (Abascal, Ezkurdia, et al. 2015).

Comparisons across protein families show that common functional effects arising from changes in domain combinations include: modifying substrate binding, creating bi-functional enzymes and functioning in a new context (Bashton & Chothia 2007). An example of how MDA changes can lead to changes in function and binding modes for the TPP superfamily is shown in **Figure 3** (Vogel & Pleiss 2014).

Enzymatic promiscuity (catalysing multiple distinct reactions from the same catalytic site) is thought to facilitate evolution of novel functions. Domain insertions can act to drive the evolution of new functions such as in the large structurally and functionally diverse HADSF enzyme superfamily(Huang et al. 2015), here different levels of promiscuity are found for different members depending on the presence or absence of a CAP domain. Those enzymes with an inserted CAP domain show wider substrate promiscuity than those without this domain. There are several possible mechanisms through which the inserted CAP domain increases promiscuity, including the juxtaposition of its loops regions close to the active site and interacting with the substrate (Huang et al. 2015). High promiscuity is also found for many members of the Alkaline Phosphatase (AP) superfamily, which achieve this through their large active sites and large polar surfaces within these sites, allowing them to provide an optimal electrostatic environment for a wide range of substrates (Pabis & Kamerlin 2016).

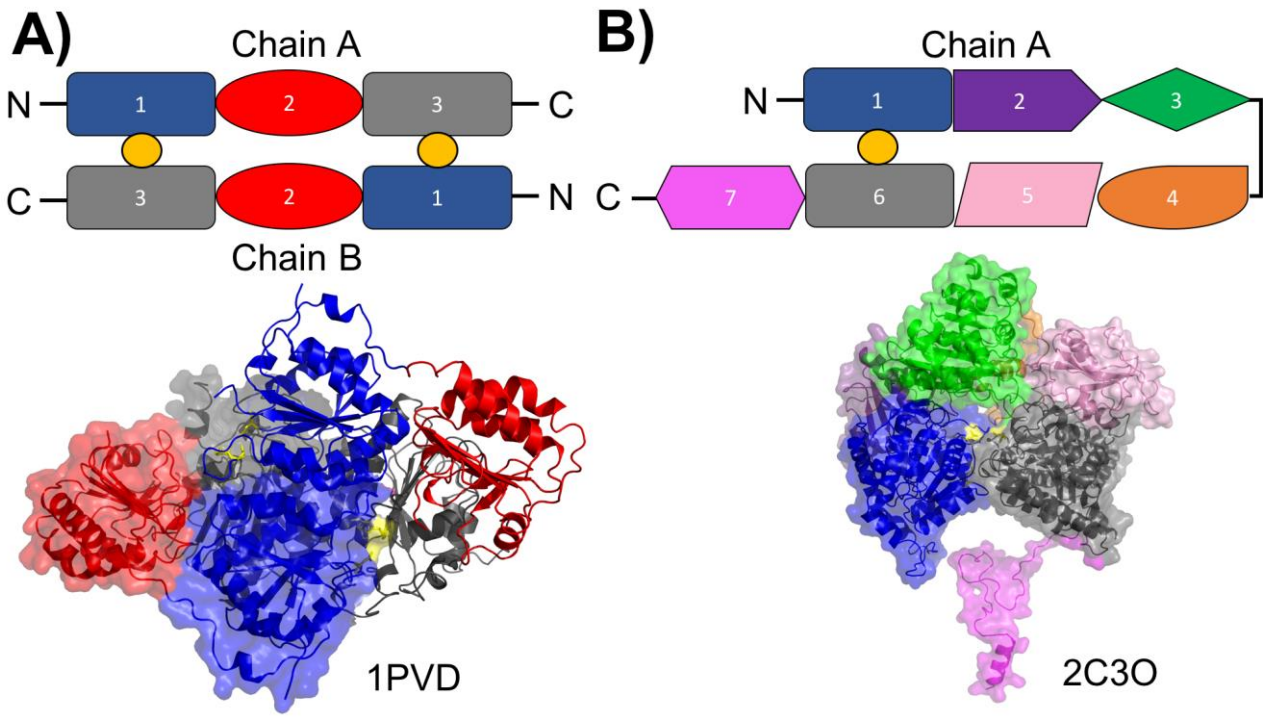

Figure 3: *The 2- and 3-dimensional **M**ulti-**D**omain **A**rchitectures (MDAs: ordered arrangement of domains) of (A) a pyruvate decarboxylase (PDB:1PVD) and (B) a pyruvate-ferredoxin oxidoreductase (PDB:2C3O) (Both proteins are homodimers, but only one monomer is shown for 2C3O (B)). The shape of the domain in the 2D representation indicates the type of CATH superfamily with individual domains having different colours. Both proteins (1PVD and 2C3O) have two CATH TPP (3.40.50.970) superfamily domains (rounded rectangles) corresponding to numbers 1 and 3 for (A) and 1 and 6 for (B). Although the proteins are homologous (through their TPP superfamily domains) they have different overall MDAs resulting in alterations to the active sites. In (A), the active sites (depicted as yellow circles / sticks) are formed between domains in different monomers, whereas in (B), the active sites are between domains in the same monomer.*

Robustness is an important factor for mediating biological innovations (Wagner 2012). An example of robustness in signalling with respect to changes in MDAs can be found in the Yeast mating pathway response, where many MDA alterations of a key kinase component of this pathway (ste5) could be engineered whilst maintaining signalling (although mainly with reduced efficiency)(Lai et al. 2015). (Robustness of network rewiring was also shown for (Sato et al. 2012)). The plethora of domain combinations in the kinase families is well known and has been catalogued in a recent review (Rakshambikai et al. 2015), highlighting the characteristic set of domain partners associated with each kinase family. However, a subset of kinase domains were found to have domain partners atypical of sometimes their own, or even any other kinase families (even though these kinase domain amino acid sequence matched well to a specific family), which could have consequences for network rewiring or cross-talk between pathways (Rakshambikai et al. 2015).

Other examples are emerging of even more intriguing functional effects from novel domain combinations. For example binding to epigenetic histone modifications is accomplished by a small set of histone 'reader' domains. The reader domains can bind in tandem to multiple histone modifications at the same time, providing combinatorial readout, with increased binding affinity and specificity (Su & Denu 2015).

Resources such as Gene3D, SUPERFAMILY, Pfam and other domain based families in

InterPro (Mitchell et al. 2015) are typically used to construct MDAs and it is reasonable to expect that the structural and functional coherence of CATH-FunFams (Das et al. 2014) will facilitate studies of emergent functions. For example, using CATH FunFams, we can see key functional innovations afforded by both intra domain-innovation (i.e. emergence of novel FunFams) and novel domain combinations in Metazoa (**Figure 4**).

Numerous studies have demonstrated the importance of domain architecture evolution for the Metazoan lineage (Ekman et al. 2007). Recently the particular importance of membrane protein diversification by extensive fusions with soluble domains (Nam et al. 2015) has been observed, with many of the added soluble domains located on the extracellular side and involved in cell communication processes. Another study suggested (although with caveats e.g. (Louis et al. 2012)) that emergence of the ancestral Chordate was accompanied by an exceptionally extensive repertoire of MDAs (Huang et al. 2014). Domain combination / function based analysis suggests channel regulators to have been an important driver of animal evolution (Linkeviciute et al. 2015). The extracellular matrix (ECM) is a fundamental Metazoan innovation and it has been shown that sampling alternative domain combinations has been important in its evolution (Cromar et al. 2014).

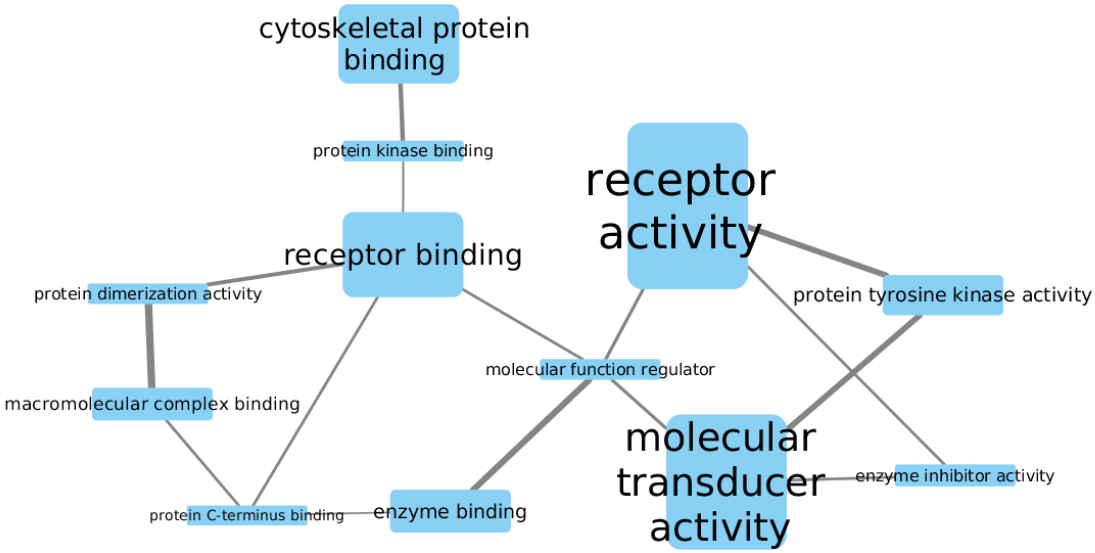

Figure 4. *Enriched intracellular-signalling GO functions in Metazoa, (comparing domains in Metazoa versus other Eukaryotes and obtaining GO terms from their CATH FunFams* (Das et al. 2014)*) (node size relates to significance of enrichment). Links between GO terms are made by analysing unique Metazoan domain superfamily pairings, obtaining the functions of each domain (from GO terms in their CATH FunFams) and then linking these functions. The thickness of links shows the relative enrichment in Metazoa relative to other eukaryotes. In order to restrict the network size, GO terms that were very general or highly specific were removed, redundant GO terms were removed and only the top 12 enriched terms were retained.*

Changes in domains and MDAs in Cancer Genomes

New technologies are providing important insights into cancer genome evolution. In cancer, there are many examples of mutations affecting individual domain activities, with a prominent example being the DNA-binding domain of p53 which frequently contains point mutations that disrupt folding and DNA binding (Selivanova & Wiman 2007). In fact, domain centric analyses

offer many advantages in understanding mutations in cancer, such as identifying mutational hotspots (Yang et al. 2015) and facilitating drug target discovery (Shi et al. 2015). Furthermore, the role of chimeric genes in cancer having unique MDAs generated through gene fusions is starting to be better understood (Mittal & McDonald 2015). For example, the FGFR3-TACC3 fusion protein (combining the FGFR kinase domain to the TACC3 tubulin binding domain) results in a protein with a constitutively active kinase localising to mitotic spindle poles leading to various chromosomal abnormalities (Singh et al. 2012). There are many examples of changes in kinase activities on gene fusion, such as loss of auto-inhibitory domains leading to constitutive activation (Stransky et al. 2014). The enrichment of certain domain combinations (Ortiz de Mend??bil et al. 2009) is helping to predict the oncogenicity of gene fusions (Shugay et al. 2013) and RNA-seq studies have revealed a major class of fusions where the novel domain combination lacks an activation domain present in one of the pre-fusion genes, possibly leading to dominant negative effects (Frenkel-Morgenstern & Valencia 2012).

Conclusions

Domains provide the key functional building blocks of proteins and many show great plasticity for functional innovation. A second level of innovation involves recombining these domains in different ways, leading to novel proteins with new molecular functions or which rewire networks or participate in different processes. Interestingly the two processes of domain change and domain combination innovations are linked. As noted above novel domain-domain interfaces can promote promiscuity and thereby facilitate the emergence of novel functions. Other examples show how binding regions of a domain can change and thereby pull the protein into new biological processes by new protein interactions (Anderson et al. 2016). Furthermore, changing the partners of a domain can open up new 'molecular environments', leading to functional innovations of the domain. For example, the Tudor domains have split into distinct functional subgroups as they have been pulled into different systems through changes in domain combinations (Jin et al. 2009).

Greater understanding of domain structures and domain combinations will have important implications for many areas of research, including cancer informatics (Yang et al. 2015; Frenkel-Morgenstern & Valencia 2012), protein/network engineering (Wang et al. 2013; Lim 2010) and automatic functional annotations (Das et al. 2014).

Acknowledgements

This work was funded to JGL by the BBSRC [BB/L002817/1] to ND by the Wellcome Trust [104960/Z/14/Z] and IS by the BBSRC [BB/K020013/1].

References

Abascal, F., Ezkurdia, I., et al., 2015. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLoS Computational Biology*, 11(6).

Abascal, F., Tress, M.L. & Valencia, A., 2015. The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome Biology*, 7(6), pp.1392–1403. Available at: http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evv076.

Abdullayev, I. et al., 2013. A reference transcriptome and inferred proteome for the salamander Notophthalmus viridescens. *Experimental Cell Research*, 319(8), pp.1187–1197.

Anderson, D.P. et al., 2016. Evolution of an ancient protein function involved in organized multicellularity in animals. *eLife*, 5, pp.1–21. Available at: http://elifesciences.org/lookup/doi/10.7554/eLife.10147.

Andreeva, A. et al., 2014. SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Research*, 42(D1).

Anon, 2010. *Introduction to Marine Genomics*, Springer Science & Business Media. Available at: https://books.google.com/books?hl=en&lr=&id=broDq7HImf8C&pgis=1 [Accessed March 1, 2016].

Bashton, M. & Chothia, C., 2007. The generation of new protein functions by the combination of domains. *Structure*, 15(1), pp.85–99. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=17223535.

Bellay, J. et al., 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome biology*, 12(2), p.R14. Available at: http://genomebiology.com/2011/12/2/R14.

Bitard-Feildel, T. et al., 2015. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis." *Biochimie*, 119, pp.244–253.

Bornberg-Bauer, E. & Albà, M.M., 2013. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, 23(3), pp.459–466.

Bornberg-Bauer, E., Schmitz, J. & Heberlein, M., 2015. Emergence of de novo proteins from "dark genomic matter" by "grow slow and moult". *Biochemical Society transactions*, 43(5), pp.867–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26517896 [Accessed February 24, 2016].

Brown, S.D. & Babbitt, P.C., 2014. New Insights about Enzyme Evolution from Large-Scale Studies of Sequence and Structure Relationships. *The Journal of biological chemistry*, pp.0–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25210038.

Bryan, P.N. & Orban, J., 2010. Proteins that switch folds. *Current Opinion in Structural Biology*, 20(4), pp.482–488.

Buljan, M. & Bateman, A., 2009. The evolution of protein domain families. *Biochemical Society transactions*, 37(Pt 4), pp.751–755.

Buljan, M., Frankish, A. & Bateman, A., 2010. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biology*, 11(7), p.R74.

Cohen-Gihon, I. et al., 2011. Evolution of domain promiscuity in eukaryotic genomes--a perspective from the inferred ancestral domain architectures. *Molecular bioSystems*, 7(3), pp.784–792.

Cromar, G. et al., 2014. New tricks for "old" domains: how novel architectures and promiscuous hubs contributed to the organization and evolution of the ECM. *Genome biology and evolution*, 6(10), pp.2897–917. Available at: http://gbe.oxfordjournals.org/content/6/10/2897.short [Accessed December 1, 2015].

Das, S. et al., 2014. Functional classification of CATH superfamilies: A domain-based approach for protein function annotation. *Bioinformatics*, 31(21), pp.3460–3467.

Das, S., Dawson, N.L. & Orengo, C.A., 2015. Diversity in protein domain superfamilies. *Current opinion in genetics & development*, 35, pp.40–49. Available at: http://www.sciencedirect.com/science/article/pii/S0959437X15000982 [Accessed October 26, 2015].

Dellus-Gur, E. et al., 2013. What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs. *Journal of Molecular Biology*, 425(14), pp.2609–2621. Available at: http://dx.doi.org/10.1016/j.jmb.2013.03.033.

Dessailly, B.H. et al., 2013. Functional site plasticity in domain superfamilies. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1834(5), pp.874–889. Available at: http://dx.doi.org/10.1016/j.bbapap.2013.02.042.

Echave, J., Spielman, S.J. & Wilke, C.O., 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2), pp.109–121. Available at: http://www.nature.com/doifinder/10.1038/nrg.2015.18.

Eddy, S.R., 2011. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10).

Edwards, H. & Deane, C.M., 2015. Structural Bridges through Fold Space. *PLoS computational biology*, 11(9), p.e1004466.

Ekman, D., Björklund, Å. K. & Elofsson, A., 2007. Quantification of the Elevated Rate of Domain Rearrangements in Metazoa. *Journal of Molecular Biology*, 372(5), pp.1337–1348.

Finn, R.D. et al., 2015. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44, pp.D279–D285. Available at: http://nar.oxfordjournals.org/content/early/2015/12/15/nar.gkv1344.full.

Forslund, K., Pekkari, I. & Sonnhammer, E.L., 2011. Domain architecture conservation in orthologs. *BMC Bioinformatics*, 12(1), p.326. Available at: http://www.biomedcentral.com/1471-2105/12/326.

Frenkel-Morgenstern, M. & Valencia, A., 2012. Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics*, 28(12), pp.i67–74. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3371848&tool=pmcentrez&rendertype=abstract [Accessed January 8, 2016].

Furnham, N. et al., 2012. Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Computational Biology*, 8(3).

Furnham, N. et al., 2015. Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies. *Journal of molecular biology*, 428(2), pp.253–267. Available at: http://www.sciencedirect.com/science/article/pii/S0022283615006531.

Goldman, A.D., Beatty, J.T. & Landweber, L.F., 2016. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *Journal of Molecular Evolution*, 82(1), pp.1–10. Available at: "http://dx.doi.org/10.1007/s00239-015-9722-8.

Hamada, M. et al., 2013. The complex NOD-like receptor repertoire of the coral acropora digitifera includes novel domain combinations. *Molecular Biology and Evolution*, 30(1), pp.167–176.

Hsu, C.-H. et al., 2016. Proteins with Highly Evolvable Domain Architectures Are Nonessential but Highly Retained. *Molecular biology and evolution*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26769031 [Accessed February 24, 2016].

Huang, H. et al., 2015. Panoramic view of a superfamily of phosphatases through substrate profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 112(16), pp.E1974–83. Available at: http://www.pnas.org/content/112/16/E1974.long [Accessed December 11, 2015].

Huang, S. et al., 2014. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nature communications*, 5, p.5896. Available at: http://www.nature.com/ncomms/2014/141219/ncomms6896/abs/ncomms6896.html.

Jack, B.R. et al., 2016. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLOS Biology*, 14(5), p.e1002452. Available at: http://dx.plos.org/10.1371/journal.pbio.1002452.

Jacob, F., 1977. Evolution and tinkering. *Science (New York, N.Y.)*, 196(4295), pp.1161–1166.

Jiang, Y. et al., 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *arXiv preprint arXiv:1601.00891*.

Jin, J. et al., 2009. Eukaryotic protein domains as functional units of cellular evolution. *Science signaling*, 2(November), p.ra76.

Lai, A., Sato, P.M. & Peisajovich, S.G., 2015. Evolution of synthetic signaling scaffolds by recombination of modular protein domains. *ACS synthetic biology*, 4(6), pp.714–722. Available at: http://pubs.acs.org/doi/abs/10.1021/sb5003482\npapers3://publication/doi/10.1021/sb5003482.

Lam, S.D. et al., 2015. Gene3D: expanding the utility of domain assignments. *Nucleic acids research*, 44(D1), pp.D404–409. Available at: http://nar.oxfordjournals.org/content/44/D1/D404.abstract?etoc.

Leonard, G. & Richards, T. a, 2012. Genome-scale comparative analysis of gene fusions, genefissions, and the fungal tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52), pp.21402–21407. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84871827466&partnerID=40&md5=0d7bb0eb2aaf15313245237704e20aff.

Light, S. et al., 2013. The impact of splicing on protein domain architecture. *Current opinion in structural biology*, 23(3), pp.451–8. Available at: http://www.sciencedirect.com/science/article/pii/S0959440X13000432 [Accessed December 11, 2015].

Lim, W.A., 2010. Designing customized cell signalling circuits. *Nature Reviews Molecular Cell Biology*, 11(6), pp.393–403. Available at: http://www.nature.com/doifinder/10.1038/nrm2904\npapers3://publication/doi/10.1038/nrm2904.

Linkeviciute, V. et al., 2015. Function-selective domain architecture plasticity potentials in eukaryotic genome evolution. *Biochimie*, 119, pp.269–277. Available at: http://www.sciencedirect.com/science/article/pii/S0300908415001376 [Accessed December 11, 2015].

Louis, A., Roest Crollius, H.R. & Robinson-Rechavi, M., 2012. How much does the amphioxus genome represent the ancestor of chordates? *Briefings in Functional Genomics*, 11(2), pp.89–95.

Marsh, J. a & Teichmann, S. a, 2010. How do proteins gain new domains? *Genome biology*, 11(7), p.126.

Mitchell, A. et al., 2015. The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Research*, 43(D1), pp.D213–D221.

Mittal, V.K. & McDonald, J.F., 2015. Integrated sequence and expression analysis of ovarian cancer structural variants underscores the importance of gene fusion regulation. *BMC medical genomics*, 8, p.40. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4504069&tool=pmcentrez&rendertype=abstract [Accessed January 6, 2016].

Murzin, A.G. et al., 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), pp.536–540.

Nam, H.-J. et al., 2015. Metazoans evolved by taking domains from soluble proteins to expand intercellular communication network. *Scientific Reports*, 5, p.9576. Available at:

http://www.nature.com/doifinder/10.1038/srep09576.

Oates, M.E. et al., 2015. The SUPERFAMILY 1.75 database in 2014: A doubling of data. *Nucleic Acids Research*, 43(D1), pp.D227–D233.

Ortiz de Mend??bil, I., Vizmanos, J.L. & Novo, F.J., 2009. Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. *PLoS ONE*, 4(3).

Pabis, A. & Kamerlin, S.C.L., 2016. Promiscuity and electrostatic flexibility in the alkaline phosphatase superfamily. *Current Opinion in Structural Biology*, 37, pp.14–21. Available at: http://dx.doi.org/10.1016/j.sbi.2015.11.008.

Philip, A.F., Kumauchi, M. & Hoff, W.D., 2010. Robustness and evolvability in the functional anatomy of a PER-ARNT-SIM (PAS) domain. *Proceedings of the National Academy of Sciences of the United States of America*, 107(42), pp.17986–17991.

Prakash, A. & Bateman, A., 2015. Domain atrophy creates rare cases of functional partial protein domains. *Genome Biology*, 16(1), pp.1–15. Available at: http://genomebiology.com/2015/16/1/88.

Radivojac, P. et al., 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3), pp.221–227.

Rakshambikai, R. et al., 2015. Typical and atypical domain combinations in human protein kinases: functions, disease causing mutations and conservation in other primates. *RSC Adv.*, 5(32), pp.25132–25148. Available at: http://xlink.rsc.org/?DOI=C4RA11685B.

Sardar, A.J. et al., 2014. The evolution of human cells in terms of protein innovation. *Molecular Biology and Evolution*, 31(6), pp.1364–1374.

Sato, P.M. et al., 2012. The Robustness of a Signaling Complex to Mutations that Alter Interaction Specificities Facilitates Network Evolution. *PLoS Biology*, 2012(12), p.e1002012. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=25490747&retmode=ref&cmd=prlinks\nfile:///Users/etromer/Documents/Papers2/Articles/2014/Sato/PLoS_Biology_2014_Sato.pdf\npapers2://publication/doi/10.1371/journal.pbio.1002012.

Scaiewicz, A. & Levitt, M., 2015. The language of the protein universe. *Current opinion in genetics & development*, 35, pp.50–56. Available at: http://www.sciencedirect.com/science/article/pii/S0959437X15000933 [Accessed November 10, 2015].

Selivanova, G. & Wiman, K.G., 2007. Reactivation of mutant p53: molecular mechanisms and therapeutic potential. *Oncogene*, 26(15), pp.2243–2254.

Shi, J. et al., 2015. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nature biotechnology*, 33(April), pp.1–10. Available at: http://www.nature.com/nbt/journal/v33/n6/full/nbt.3235.html#ref1\nhttp://www.ncbi.nlm.nih.gov/pubmed/25961408.

Shugay, M. et al., 2013. Oncofuse: A computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics*, 29(20), pp.2539–2546.

Sillitoe, I. et al., 2015. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1), pp.D376–D381.

Singh, D. et al., 2012. Transforming Fusions of FGFR and TACC Genes in Human Glioblastoma. *Science*, 337(6099), pp.1231–1235. Available at: http://www.sciencemag.org/content/337/6099/1231.

Stransky, N. et al., 2014. The landscape of kinase fusions in cancer. *Nature communications*, 5,

p.4846. Available at: http://www.nature.com/ncomms/2014/140910/ncomms5846/full/ncomms5846.html.

Studer, R. a et al., 2014. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences of the United States of America*, 111(6), pp.2223–8. Available at: http://www.pnas.org/content/early/2014/01/23/1310811111.abstract.html?etoc.

Su, Z. & Denu, J.M., 2015. Reading the Combinatorial Histone Language.

Todd, A.E., Orengo, C. a & Thornton, J.M., 2001. Evolution of function in protein superfamilies, from a structural perspective. *Journal of molecular biology*, 307(4), pp.1113–1143. Available at: http://www.sciencedirect.com/science/article/pii/S0022283601945139.

Tompa, P. & Fersht, A., 2009. *Structure and function of intrinsically disordered proteins*, Chapman and Hall/CRC. Available at: https://books.google.co.uk/books?hl=en&lr=&id=GzuxFYrzfd4C&oi=fnd&pg=PP1&dq=Structure+and+Function+of+Intrinsically+Disordered+Proteins+pdf&ots=qUWHKkvXNx&sig=GD_tXDNBZG-hoornzqWsA14ePnM [Accessed February 23, 2016].

Toth-Petroczy, A. & Tawfik, D.S., 2014. The robustness and innovability of protein folds. *Current Opinion in Structural Biology*, 26(1), pp.131–138.

Triant, D.A. & Pearson, W.R., 2015. Most partial domains in proteins are alignment and annotation artifacts. *Genome biology*, 16(1), p.99. Available at: http://genomebiology.com/2015/16/1/99.

Vogel, C. et al., 2004. Supra-domains: Evolutionary Units Larger than Single Protein Domains. *Journal of Molecular Biology*, 336(3), pp.809–823.

Vogel, C. & Pleiss, J., 2014. The modular structure of ThDP-dependent enzymes. *Proteins*, 82(10), pp.2523–37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24888727 [Accessed December 11, 2015].

Wagner, A., 2012. The role of robustness in phenotypic adaptation and innovation. *Proceedings of the Royal Society B: Biological Sciences*, 279(1732), pp.1249–1258. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3282381&tool=pmcentrez&rendertype=abstract.

Wagner, G.P., Pavlicev, M. & Cheverud, J.M., 2007. The road to modularity. *Nature reviews. Genetics*, 8(12), pp.921–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18007649.

Wang, B. et al., 2013. Rewiring cell signalling through chimaeric regulatory protein engineering. *Biochemical Society transactions*, 41(5), pp.1195–200. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3782828&tool=pmcentrez&rendertype=abstract.

Weiner, J., Beaussart, F. & Bornberg-Bauer, E., 2006. Domain deletions and substitutions in the modular protein evolution. *FEBS Journal*, 273(9), pp.2037–2047.

Wolf, Y.I. & Koonin, E. V., 2013. Genome reduction as the dominant mode of evolution. *BioEssays*, 35(9), pp.829–837.

Wright, P.E. & Dyson, H.J., 2014. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1), pp.18–29. Available at: http://dx.doi.org/10.1038/nrm3920.

Yang, F. et al., 2015. Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. *PLoS Computational Biology*, 11(3).

Yang, X. et al., 2016. Widespread Expansion of Protein Interaction Capabilities by Alternative

Splicing. *Cell*, 164(4), pp.805–817. Available at: http://www.cell.com/article/S0092867416300435/fulltext.

Zhang, D. et al., 2014. Resilience of biochemical activity in protein domains in the face of structural divergence. *Current Opinion in Structural Biology*, 26(1), pp.92–103. Available at: http://dx.doi.org/10.1016/j.sbi.2014.05.008.

Zmasek, C.M. & Godzik, A., 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biology*, 12(1), p.R4. Available at: http://genomebiology.com/2011/12/1/R4\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3091302&tool=pmcentrez&rendertype=abstract.