Genome Biology

RESEARCH HIGHLIGHT

How do proteins gain new domains?

Joseph A Marsh and Sarah A Teichmann*

Abstract

A study of the contributions of different mechanisms of domain gain in animal proteins suggests that gene fusion is likely to be most frequent.

Domains are evolutionarily conserved regions of proteins with generally independent structural and functional properties. Although only a fairly limited set of domains has been created during evolution, combining these domains in different ways has led to the huge number of observed protein domain architectures. These multidomain proteins have diverse functions that rely on the collective properties of their component domains. Therefore, a key to understanding the evolution of proteins is to understand how multidomain proteins gain, lose and rearrange domains. A considerable body of literature has been dedicated to extrapolating these mechanisms from amino acid sequence and domain architecture information [1-5]. In a study in this issue of *Genome Biology*, Buljan *et al*. [6] have addressed the question from a new perspective - by investigating the relative contributions of different molecular genetic mechanisms for domain acquisition to the evolution of animal proteins, inferred from gene structure at the nucleotide level.

The availability of a large number of fully sequenced genomes in recent years has facilitated significant insight into the evolution of domain architectures in multidomain proteins. The tendency for proteins to exist in multidomain combinations has been found to differ greatly between different branches of the evolutionary tree, with eukaryotes generally having a greater proportion of multidomain proteins [1]. Animal proteins are particularly interesting, as the creation of multidomain proteins and the rate of domain rearrangements appear to have substantially increased in the recent metazoan lineage [2]. Different protein-domain families have widely varying propensities to combine with other domains: most will combine with very few other domains, whereas some will form a large number of combinations [1]. Most evolutionary changes to multidomain protein architectures occur at the amino and carboxyl termini in the form of insertions of new domains, domain repetitions and domain deletions [3,4]. Recent modeling at the protein-sequence level suggests that the evolution of most protein-domain architectures can be explained by a series of simple steps, and that complex rearrangements are rare [5].

Mechanisms for domain acquisition

Proteins can acquire new domains by various mechanisms. Gene fusion, in which two adjacent genes become joined, is a major mechanism for multidomain protein formation in bacteria [7]. However, the mechanisms for domain gain in eukaryotes are more varied, primarily because of their complex exon-intron gene structures. Although gene fusion is also important in eukaryotes, it typically does not involve the direct joining of exons from adjacent genes. Instead, splicing patterns are modified so that a fused gene is transcribed from the still separated exons (Figure 1a). Interestingly, the rate of gene fusion appears to be considerably greater than the opposite process, gene fission, in which a single gene splits into two [5].

A different mechanism for domain gain involves the extension of an exon into a noncoding region (Figure 1b). One might presume this mechanism to be extremely rare, given that expression of a previously noncoding sequence would seem unlikely to result in a functional polypeptide. Buljan *et al*. have specifically addressed this mechanism, as we discuss later.

Other mechanisms for protein domain gain involve recombination. For example, exons from two different genes could be directly joined (Figure 1c). Alternatively, exons from one gene could be inserted into the introns of another (Figure 1d). Intronic recombination is often referred to as exon shuffling, and has been speculated to be one of the main drivers behind the diversity of domain architectures in complex eukaryotes [8]. An important role for intron recombination in domain rearrangements is supported by the observations that there are significant correlations between domain boundaries and exon boundaries, and that most of the exons that correspond to domains are surrounded by introns of symmetric

*Correspondence: sat@mrc-lmb.cam.ac.uk
MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

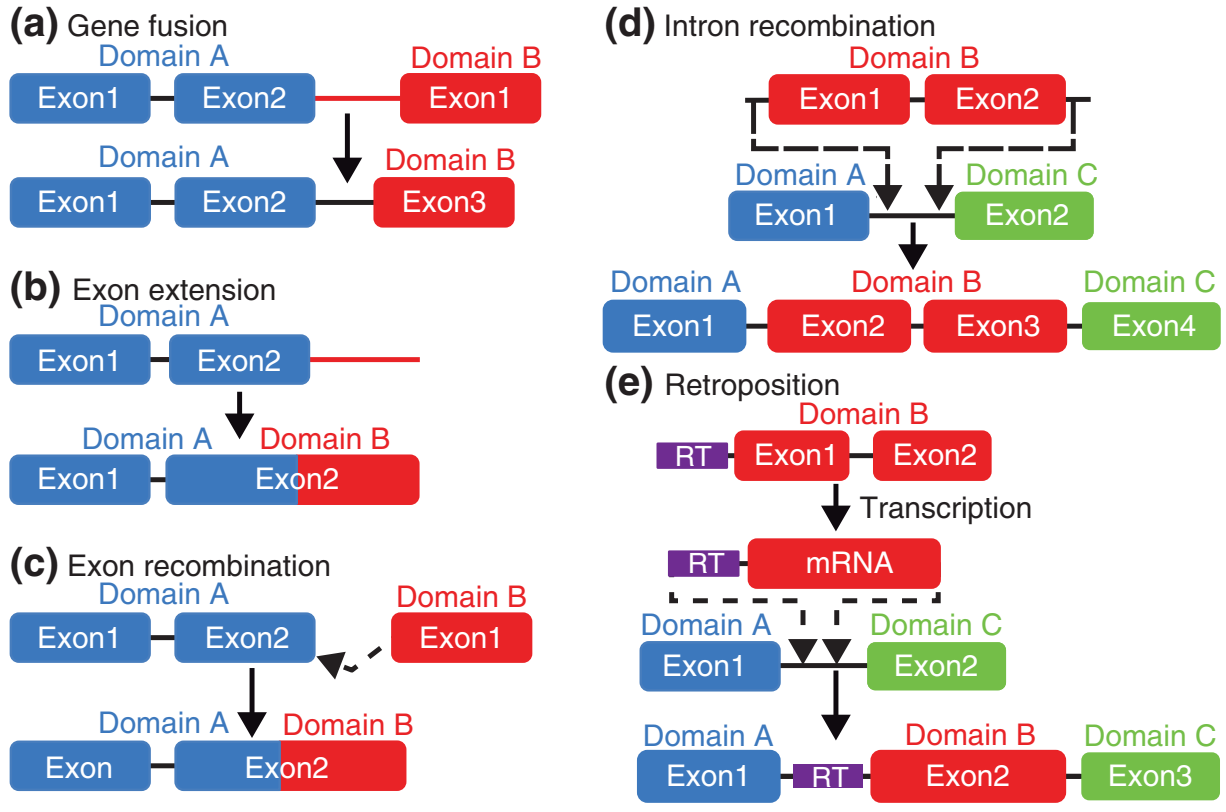

Figure 1. Possible mechanisms for the gain of protein domains. Colored blocks represent exons, with red, blue and green indicating exons coding for different domains. Solid black lines represent introns and red lines indicate intergenic regions. **(a)** Gene fusion. The noncoding region between two genes is modified so that the exons of the first gene become spliced with the second. **(b)** Exon extension. The noncoding region following an exon becomes part of the exon and codes for a new domain. **(c)** Exon recombination. The exons of two genes become directly joined. **(d)** Intron recombination. An exon from one gene is inserted into the intron of another. **(e)** Retroposition. A retrotransposon sequence (RT, purple) mediates the copying of itself and a neighboring gene region via an mRNA intermediate, followed by insertion into another gene.

phase (that is, introns are inserted at the same positions with respect to codon triplets) [9].

Retrotransposons are genetic elements that can replicate and insert themselves at other genomic locations. This provides another possible mechanism for protein domain gain, as retrotransposons can also copy regions of genes and insert them into other genes (Figure 1e). Notably, because retroposition occurs via an mRNA intermediate, an inserted region will lack the introns of the gene from which it originated.

Assessing the relative contributions of domain-gain mechanisms

Although the actual physical events behind most domain gains may be more complex than presented in Figure 1, these mechanisms provide a simple framework by which the majority of protein domain gains can be explained. However, despite the recent work on multidomain protein evolution at the amino acid level, there has been little investigation of the extent to which the different molecular genetic mechanisms have contributed to the current diversity of multidomain protein architectures in complex eukaryotes. This is the question that Buljan *et al.* [6] have set out to address.

The authors started by compiling a set of putative domain-gain events. These were identified by examining the domain assignments and phylogenetic relationships between genes from a large number of fully sequenced genomes. As previous work has shown that the process of identifying evolutionary changes in domain architectures can be sensitive to erroneous annotations [3], the authors used very stringent criteria in their selection process to ensure that the identified gains were likely to be true domain-gain events and not domain losses or artifacts of the genome or domain annotation procedures. Thus, although this procedure is likely to miss some true gains, the final set, containing 330 high-confidence domain-gain events, should include very few false positives.

The key to assessing the relative contributions of different domain-gain mechanisms is the fact that

different mechanisms should leave distinct genomic traces. For example, a domain gained from retroposition is likely to have only a single exon as the retrotransposon replicates via a transcribed mRNA intermediate. Thus, gained domains containing multiple exons are unlikely to have been acquired via retroposition. Other mechanisms, including gene fusion and exon recombination, are much more likely to occur at protein termini, whereas intron recombination can only occur in the middle of a protein. The location of the gained domain can thereby be used to infer by what mechanisms the domain gain was likely to have occurred. Finally, for all gained domains, the authors searched for homologs within the same genomes to identify potential 'donor' genes. This provides information on whether gene duplication preceded domain gains and can identify potential source genes for retroposition.

A primary finding of this study [6] was that most domain gains (71% of the total) occurred at the amino or carboxyl termini of proteins, and that most of these gains involved multiple exons. Gene fusion is the only plausible mechanism that can account for these 32% of gains that occur at termini and involve multiple exons. In addition, gene fusion is likely to have caused many of the other 39% of gains that occurred at termini, although, in these cases, other mechanisms cannot be excluded. These results strongly suggest that gene fusion is the most important mechanism for domain gain in animals. Of course, fusion can only occur between genes that are adjacent on the chromosome. The authors found no evidence that any of the fused genes existed separately in adjacent, non-fused forms, and so an additional mechanism would be required to juxtapose the genes before fusion. In at least 80% of domain-gain events, there was evidence for duplication preceding the domain gain of either the donor gene or the gene that acquired the domain. In addition, in cases where a donor gene could be identified in the same genome, it was located on the same chromosome as the domain gain in a significant fraction of these cases. This strongly suggests nonallelic homologous recombination as the likely mechanism for bringing separate genes together, as it favors recombination on the same chromosome.

Although recombination between introns has been speculated to be one of the main mechanisms behind the diverse domain rearrangements observed in complex eukaryotes [8], it seems to have made a fairly limited contribution to the domain-gain events studied by Buljan *et al.* [6]. Only 10% of the gained domains were both internally located and surrounded by introns of symmetric phase, which would make their gain likely to have occurred by intron recombination. Thus, although it has probably played a very important role in the evolution of some multidomain proteins, intron recombination has contributed to far fewer domain gains than has gene fusion.

Gained domains that were encoded by single exons and for which potential donor genes could be identified are likely candidates for retroposition. Only a few gains fit these criteria, and manual inspection revealed only a single case in which a retrotransposon sequence was present in the donor gene. Thus, the authors [6] suggest that retroposition underlies only a small fraction of domain gains in animal proteins. However, they do note a high percentage of single-exon domain gains in insects, which hints that retroposition may have played different roles in different lineages.

A very interesting finding from this study relates to the frequency of intrinsically disordered regions in the gained domains. Intrinsically disordered regions of proteins lack stable folded structure, and have recently garnered significant attention because of their numerous important biological functions and their association with various human diseases [10]. Interestingly, the authors noted that the fraction of residues predicted to be intrinsically disordered was significantly greater in gained domains than in other domains. In particular, those domains encoded by exon extensions showed a dramatic enrichment in disorder. This suggests an origin for these disordered regions from previously noncoding sequences that have become exonized. Thus, this study has important implications for both understanding the origin of intrinsically disordered protein sequences and for helping to explain the preponderance of proteins in complex eukaryotes that possess intrinsically disordered regions. Figure 2 shows a hypothetical example of a protein with multiple folded domains gaining an intrinsically disordered region at its carboxyl terminus via an exon extension.

Inferring evolutionary mechanisms from genomic sequences with millions of years of divergence between them is inherently difficult and Buljan *et al.* [6] have done an admirable job of extracting the available information. However, there is still considerable work to do to improve our understanding of different domain-gain mechanisms. Evolution is complex, and it is likely that a mixture of processes contributed to many domain gains and rearrangements. For example, although gene fusion is likely to be the dominant domain-gain mechanism, the recombination that precedes it relies on regions of sequence similarity that may have originated from retrotransposon activity. Moreover, the methods for classifying domain gains from sequences are imperfect and thus frequencies given for different domain-gain mechanisms can only be considered rough estimates. Nonetheless, this study [6] provides strong support for the idea that most domain gains in animal proteins were directly mediated by gene fusion, preceded by duplication and recombination. Intron recombination and retroposition, on the other hand, appear to have been less

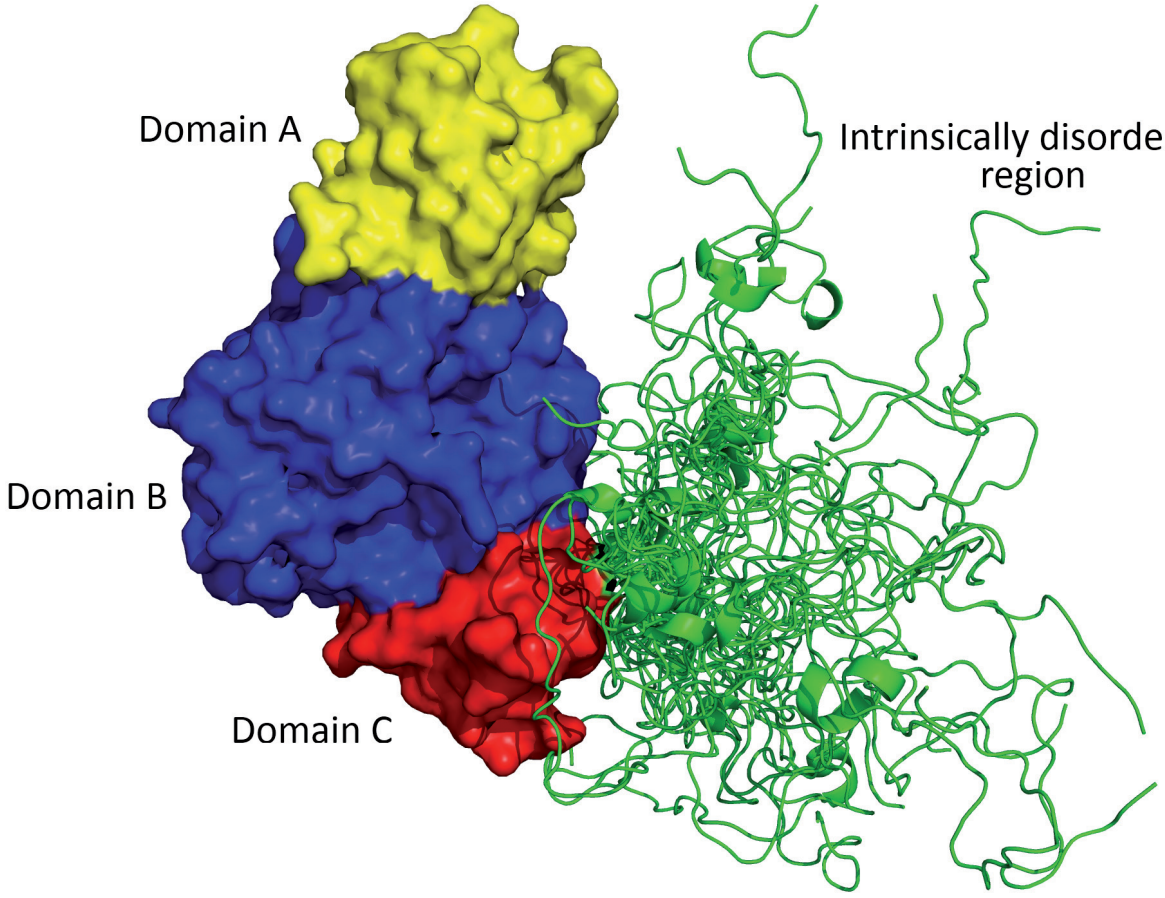

Figure 2. Hypothetical model of a multidomain protein gaining an intrinsically disordered region via a carboxy-terminal exon extension. This protein has three folded domains (based on Protein Data Bank entry 1BIB), colored yellow, blue and red, and a 40-residue disordered extension at its carboxyl terminus, colored green. The folded domains are shown as a surface representation, and the disordered region is shown as an ensemble model with multiple distinct structures representing its substantial conformational heterogeneity.

important in recent evolutionary history. Because of the tremendous recent advances in next-generation sequencing technologies, the number of fully sequenced genomes will vastly increase in the relatively near future. This will allow the molecular genetic mechanisms of multidomain protein evolution to be studied in much more detail.

Acknowledgements
JM is supported by an EMBO Long-Term Fellowship.

Published: 15 July 2010

References

1. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *J Mol Biol* 2001, **310:**311-325.
2. Ekman D, Björklund AK, Elofsson A: **Quantification of the elevated rate of domain rearrangements in metazoa.** *J Mol Biol* 2007, **372:**1337-1348.
3. Weiner J, Beaussart F, Bornberg-Bauer E: **Domain deletions and substitutions in the modular protein evolution.** *FEBS J* 2006, **273:**2037-2047.
4. Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A: **Domain rearrangements in protein evolution.** *J Mol Biol* 2005, **353:**911-923.
5. Fong JH, Geer LY, Panchenko AR, Bryant SH: **Modeling the evolution of protein domain architectures using maximum parsimony.** *J Mol Biol* 2007, **366:**307-315.
6. Buljan M, Frankish A, Bateman A: **Quantifying the mechanisms of domain gain in animal proteins.** *Genome Biol* 2010, **11:**R74.
7. Pasek S, Risler J, Brézellec P: **Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins.** *Bioinformatics* 2006, **22:**1418-1423.
8. Patthy L: **Genome evolution and the evolution of exon-shuffling - a review.** *Gene* 1999, **238:**103-114.
9. Liu M, Grigoriev A: **Protein domains correlate strongly with exons in multiple eukaryotic genomes - evidence of exon shuffling?** *Trends Genet* 2004, **20:**399-403.
10. Dyson HJ, Wright PE: **Intrinsically unstructured proteins and their functions.** *Nat Rev Mol Cell Biol* 2005, **6:**197-208.

doi:10.1186/gb-2010-11-7-126
Cite this article as: Marsh JA, Teichmann SA: **How do proteins gain new domains?** *Genome Biology* 2010, **11:**126.