

Chapter 3

Mechanisms of domain gain in animal proteins

3.1 Introduction

In the previous chapter, I discussed general trends in the evolution of animal protein domain architectures. However, I also showed there that reported domain gain and loss events strongly depend on their initially assumed relative frequencies. Hence, to be able to investigate signatures of the causative mechanisms for these changes it is necessary first to compose a set of clear, confident events. The creation of more complex domain architectures is crucial for the evolution of complexity in animals and this chapter focuses on the mechanisms for insertion of novel domains into ancestral proteins. Novel domain combinations are a basis for the invention of original protein functions and lay at the heart of evolution of species-specific traits (Kawashima et al., 2009).

Eukaryotic domain architectures are far more complex than prokaryotic ones, and it is believed that the underlying reason for this is a greater choice of mechanisms that can create novel domain combinations (Chothia et al., 2003). The main eukaryote-specific mechanisms are intronic recombination, joining of

adjacent genes' exons preceded by intergenic splicing and retroposition. I will first introduce here the concept of 'exon shuffling through intronic recombination', which was widely discussed as a powerful means for evolution of novel domain architectures, and then elaborate further on other mechanisms that are assumed to be active in eukaryotic genomes and are able to cause domain gain.

It has been recognized for a long time that intronic sequences can mediate gene recombination and thereby cause exon shuffling (Gilbert, 1978). Intronic recombination can either join the termini of two different genes or insert novel exons into ancestral introns. To date, specific examples in animals have been reported for domain gains through exon insertions into introns and a term 'domain shuffling through intronic recombination' was devised to describe this phenomenon (Patthy, 1996). The extracellular function of the inserted domains indicates the importance of this mechanism for the evolution of multicellular organisms. Additionally, more recent whole-genome studies of domain shuffling have also focused on domains that are candidates for exon insertions into introns, for example; domains that are surrounded by introns of symmetrical phases (Kaessmann et al., 2002; Liu and Grigoriev, 2004; Long et al., 1995). Phase of an intron is defined by the break point in the codon next to the intron. For example, if an intron is placed after the first nucleotide in the codon, it is phase 1 intron. Analogously, if it is placed after the second nucleotide, it is phase 2, and if it is placed after all three nucleotides in the codon, it is phase 0 (Figure 3.1). When a new exon is inserted into an ancestral intron, it needs to be surrounded by introns of symmetrical phases for it to be translated in frame and not to disrupt the translation of the downstream sequence. The studies that found an excess of domains surrounded by symmetrical introns in the genomes of higher eukaryotes suggested that domain insertions into introns have had an important role in the evolution of eukaryotic proteomes. It is noteworthy that even though initial studies attributed intronic insertions solely to intronic recombination, authors of the more recent studies have also acknowledged the potential role of retroposition (which is described below) in this process.

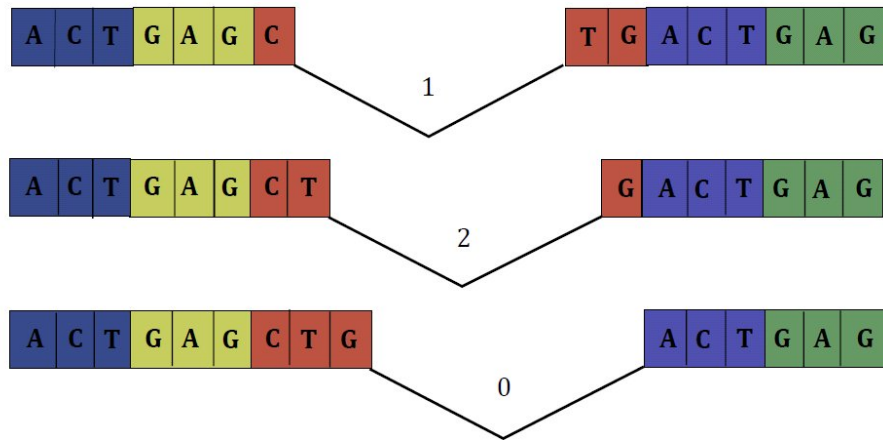


Figure 3.1. Illustration of intron phases. Phase of an intron is defined by the breakpoint in the codon adjacent to the intron.

The question of what mechanisms underlie domain gains is related to the question of what mechanisms underlie novel gene creation (Babushok et al., 2007b), (Arguello et al., 2007; Long et al., 2003). The recent increased availability of animal genome and transcriptome sequences offers a valuable resource for addressing these questions. The main genetic mechanisms that are capable of creating novel genes and also causing domain gain in animals are retroposition, gene fusion through joining of exons from adjacent genes, and DNA recombination (Arguello et al., 2007; Babushok et al., 2007b; Long et al., 2003) (Figure 3.2). Since these mechanisms can leave specific traces in the genome, it may be possible to infer the causative mechanism by inspecting the DNA sequence that encodes the gained domain. By using the retrotransposon machinery, in a process termed retroposition, a native coding sequence can be copied and inserted somewhere else in the genome. The copy is made from a processed mRNA, so sequences gained by this mechanism are usually intronless and have an origin in the same genome. This was proposed as a powerful means for domain shuffling, but the evidence for its action is still limited (Babushok et al., 2007a; Zhou et al., 2008). Recent studies observed a phenomenon where adjacent genes, or nearby genes on the same strand undergo intergenic splicing and create chimerical transcripts (Akiva et al., 2006; Magrangeas et al., 1998;

Parra et al., 2006). This suggested that if promoter and terminator sequences between the two genes were degraded during evolution then exons of the genes could be joined not only on the transcript level, but also as a novel chimeric gene. As a consequence of this, one would observe a gain of novel exon(s) at the protein termini. One example for this mechanism is the creation of the human gene Kua-UEV (Thomson et al., 2000). Recombination can aid novel gene creation by juxtaposing new gene combinations, thereby assisting exons from adjacent genes to combine. When recombination occurs between intronic sequences of two genes and joins the genes by creating a novel chimerical intron, then joining of exons from the adjacent genes is in concordance with the theory of exon shuffling through intronic recombination. Alternatively, recombination could occur between exonic sequences of two different genes (Patthy, 2008). The two main types of recombination are non-allelic homologous recombination (NAHR) (Arguello et al., 2007; Turner et al., 2008), which relies on short regions of homology, and illegitimate recombination (IR) – also known as non-homologous end joining (Arguello et al., 2007; Long et al., 2003; van Rijk and Bloemendal, 2003). IR does not require homology regions for its action, but instead can join DNA breaks with no similarity at all, or with similarity of only several nucleotides. In addition to these mechanisms, a new protein coding sequence can be gained through (i) deletion of the intervening sequence between two adjacent genes and subsequent exon fusion (Nurminsky et al., 1998); (ii) by exonisation of previously non-coding sequence (Zhang and Chasin, 2006); (iii) through insertion of viral or transposon sequences into a gene (Cordaux et al., 2006). Interestingly, direct examples for any of these mechanisms are still rare (Babushok et al., 2007a; Thomson et al., 2000).

In this chapter, I will first describe a procedure that I applied for identification of a set of confident domain gain events and the control steps I implemented to ensure that the reported gain events are not due to gene annotation errors or method bias. Next, I will describe the results of the analysis of the sequences that encode these domains. The study of signatures of possible causative mechanisms for these domain gains suggested that gene fusion through joining of exons from adjacent genes has been a dominant process leading to gains of new domains. Two other mechanisms that have been

proposed as important mediators for gains of new domains in animals - retroposition and 'exon shuffling through intronic recombination' - appear to be minor contributors. In concordance with the results in Chapter 2, I observe here that gene duplications play an important role in domain gains. Finally, several lines of evidence suggest that these domain gain events were assisted by DNA recombination, and trends in these gain events point to NAHR as a possible acting mechanism.

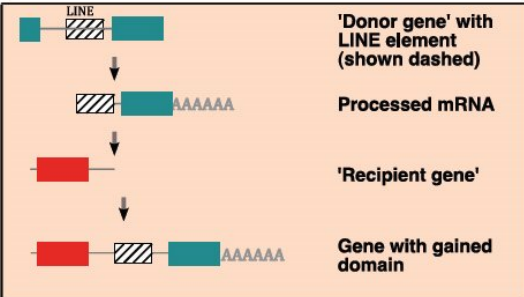
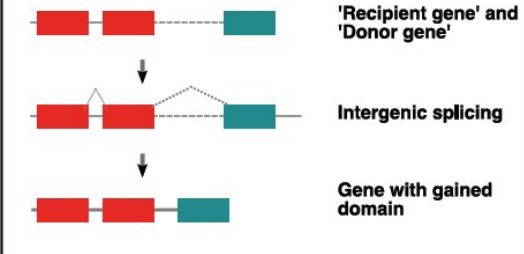

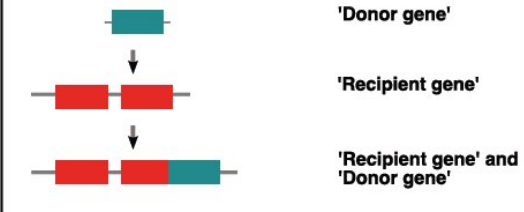
Result of domain gain	Possible causative mechanism	Position of gain	Number of exons gained
 <p>'Donor gene' with LINE element (shown dashed)</p> <p>Processed mRNA</p> <p>'Recipient gene'</p> <p>Gene with gained domain</p>	Retroposition.	Anywhere in a protein.	Only one, since the intermediate step is reverse transcription from a processed mRNA.
 <p>'Recipient gene' and 'Donor gene'</p> <p>Intergenic splicing</p> <p>Gene with gained domain</p>	Gene fusion through joining of exons from adjacent genes, possibly preceded by intergenic splicing. Also, initially non adjacent genes could have become juxtaposed by NAHR or IR.	Protein termini.	One or more.
 <p>'Donor gene'</p> <p>'Recipient gene'</p> <p>'Recipient gene' and 'Donor gene'</p>	NAHR, IR or retroposition can mediate gain of novel middle exons.	Middle of a protein.	One or more if the causative mechanism is recombination. Only one if the causative mechanism is retroposition.
 <p>'Donor gene'</p> <p>'Recipient gene'</p> <p>'Recipient gene' and 'Donor gene'</p>	NAHR or IR between exons of two separate genes are presumably the most likely causative mechanisms for exon extensions where 'donor genes' can be found.	Most likely at protein termini.	Either only an existing exon is extended, or additional exons are gained as well.

Figure 3.2: Summary of mechanisms for domain gains. This figure shows mechanisms that can lead to domain gains and the signals that can be used to detect the causative mechanism. Domain gain by retroposition is illustrated as an example where the domain is transcribed together with the upstream long interspersed nuclear element (LINE), but other means of retroposition are also possible (Babushok et al., 2007b). The list of possible mechanisms is not exhaustive and other scenarios can occur, as, for example, exonisation of previously non coding sequence or gain of a viral or transposon domain during retroelement replication.

3.2 Methods

3.2.1 Assignment of domains to proteins with refinement

Pfam domains (release 23.0) were assigned to all protein products of genes in the TreeFam database (release 6.0) using the Pfam_scan.pl software. The same procedure for refinement of domain assignments that is described in Chapter 2 was applied here; domain identifiers were replaced with clan identifiers, false domain assignments were removed and missing domain assignments were added to proteins. Methodological details of this are explained in Chapter 2.2.2.

3.2.2 Exclusion of possible false domain gain calls

Domain refinements described above added Pfam domains to proteins that shared significant similarity with annotated domain sequences but were not recognized by searching with the Pfam HMM library. However, apart from these clear cases of a lack of domain annotation, there are also cases where proteins share only moderate similarity with domain sequences and it is difficult to say whether a domain should be annotated to these proteins as well. To be able to do this analysis, a set of confident domain gains was crucial. Hence, in order to avoid false calls of domain gains, domain gain events where sequences in the same gene family shared a similarity with the gained domain but were not annotated with that domain were excluded. This included all gain events where a domain sequence had 16% or more identical amino acids aligned to any sequence in the same TreeFam family that lacked the gained domain. This threshold was justified by distribution of fractions of identical amino acids in the initially reported domain gain events (Appendix B.1). This is in agreement with the expectation that initially reported domain gain events are a mixture of true gain events and false calls caused by errors in domain annotations. A 16% sequence identity was noted as a threshold that apparently separated the majority of these events. This filtering step further reduced the chances of

erroneously calling domain gains due to a lack of sensitivity of some Pfam HMM models.

3.2.3 Parsing trees

To identify the branch points in the phylogenetic trees at which new domains were gained the TreeFam API (Ruan et al., 2008) was used. In TreeFam families each gene is represented with a single transcript. However, to be able to claim that a gene has gained a domain it was necessary to take into account protein domains present in all splice variants of the genes in the TreeFam families. The weighted parsimony algorithm (Sankoff et al., 1982) was applied on the TreeFam phylogenies, with the cost for a domain gain of 2 and the cost for a domain loss of 1. Because gains are more costly, the ones that are reported are more likely to be correct. However, only those reported gain events that occurred once in a tree - which is the rationale of the Dollo parsimony (Farris, 1977) - were taken into account. This condition removed from the set instances where domain gains were inferred several times in a gene family, and where multiple domain losses could have also explained the differences in domain architectures of present proteins. This method was applied to the 17,050 TreeFam clean trees, i.e. trees containing genes from completely sequenced animal genomes. Events that were in concordance with both algorithms were considered as likely gain events - these included 4362 gained domains.

Gain events that appeared on the leaf nodes of the trees, i.e., which had only one sequence with the gained domain, were excluded from further analysis. When a domain gain is not supported by at least two proteins, the gain is less reliable because it could also be a consequence of an incorrect gene annotation process. This left 1372 domains gained on internal nodes of the tree. Next, one representative transcript for each gain event was chosen. The approach for choosing the representative transcript was the following: the transcript had to be the one present in the TreeFam tree, a representative transcript had to have a gained domain predicted initially by the Pfam software and finally, the representative transcript had to belong to one of the following species: *Drosophila melanogaster* (fruit fly), *Xenopus tropicalis* (frog), *Danio rerio*

(zebrafish), *Gallus Gallus* (chicken), *Mus musculus* (mouse), *Rattus norvegicus* (rat) or *Homo sapiens* (human). Thus, the study included the major animal model organisms. The advantage of this is that a majority of these organisms have genomes of better quality; an exception being chicken and rat genomes. There were 653 gained domains that had representative transcripts which fulfilled all conditions. Since each representative sequence was chosen from a descendant with the genome of best quality, for all gains in the human lineage the representative sequence was a human transcript (protein). Exclusion of leaf gains and selection of representative transcripts from better quality genomes were necessary to ensure that the reported gain events were not due to gene annotation errors. Next, all instances where a sequence from the same family that lacked the gained domain was found to have diagnostic motifs for that domain, as recognized by profile comparer (Madera, 2008), were excluded, as well as the instances where a sequence without domain annotation had an amino acid stretch similar to one in the gained domain (16% or more identical amino acids, explained above). This left us with 378 gained domains in the set. Some of these domains appeared to be gained as a result of the same event that extended the ancestral gene, so the total number of domain gain events was 349. Finally, the following cases were also excluded from the analysis: the gain events for which a representative transcript was no longer in the Ensembl database, release 50 (3 cases), events for which protein sequence alignment downloaded from the TreeFam database did not clearly support domain gain (13 cases) and the cases that were later found to be most likely consequences of inconsistencies in gene annotation (3 cases). The final set had a total of 330 high confidence domain gain events (Appendix B.2). Still, sometimes the same gene has experienced more than one domain gain, and a total number of representative sequences for the 330 domain gains was 322 (Appendix B.2).

To investigate whether the set of high-confidence domain gains discriminates against any mechanism because of a small number of events, a set of medium confidence domain gain events was created. For this, the same initial set of reported gain events was taken and the applied condition was that each gain had to occur in at least one genome of better quality. Other filtering steps were omitted. Hence, gains on the leaf nodes, as well similarity of the 'gained

domain' with sequences in the same family that were not annotated with that domain were allowed. Consequently, this also increased the rate of false calls of domain gains. There were 849 gained domains in the set of medium confidence domain gain events. The flow of the procedures for obtaining of the high and medium confidence sets of gain events is illustrated in Figure 3.3 and the flow of the procedures for the analysis of these gains in Figure 3.4.

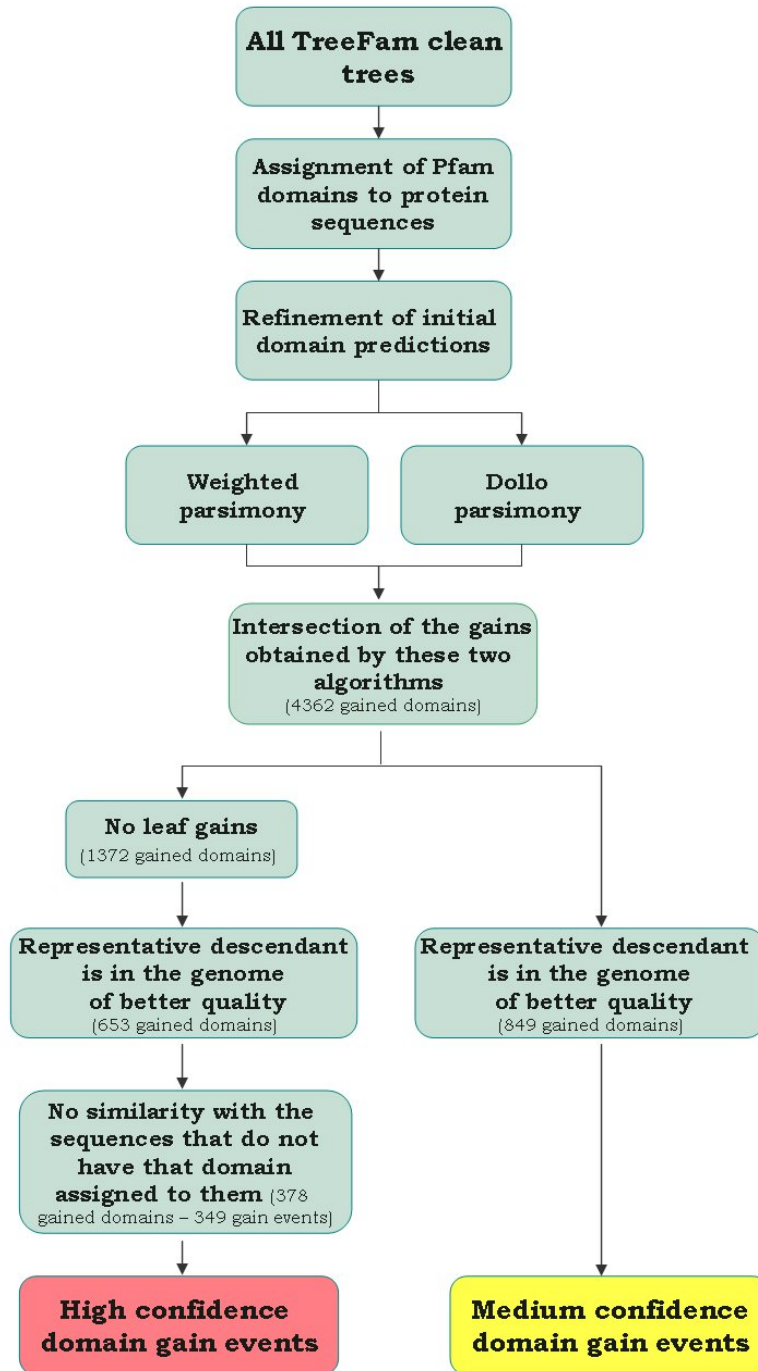


Figure 3.3: Flowchart of methods for obtaining sets of high and medium confidence domain gain. The numbers of gained domains I was left with after each filtering step are noted. In some cases more domains were gained at the same time; hence the number of gain events that we looked at for the high confidence domain gains differs from the number of gained domains.

3.2.4 Intron-exon structures of genes

The TreeFam table Map with gene structures was used to project the intron-exon boundaries and intron phases on the representative protein sequences for each domain gain event. The goal of the analysis was to investigate the type of changes that occurred on the gene level when a domain was gained; in particular whether a domain gain was the result of a gain of a new exon or extension of an already existing exon. To infer this, protein sequence alignments for each TreeFam family with a gained domain were downloaded from the TreeFam website. In order to establish whether the gained protein domain was part of a completely new exon or an extension of a pre-existing exon, the similarity in regions close to the exon boundaries was examined. If the region in the same exon close to the exon border shared partial similarity with an exon from the protein in the same family that lacked the domain, a domain gain was considered to be the result of an exon extension. The criterion for similarity was that the first or last third of the sequence outside of the domain – adjacent to the exon border - had 30% or more identical residues to one of the sequences without the inserted domain. It was required that this 'boundary' region was at least seven amino acids long. However, because of this criterion that only a short stretch of sequence similarity is enough to claim that a gained domain is coded by an extended ancestral exon, the number of extended exons is likely to be an overestimate.

3.2.5 Positions of gained domains

When a new domain was coded by the first or last coding exon the gain was called an N- or C-terminal gain, respectively. In addition, when an inserted domain was not coded by the terminal exons, it was checked whether additional exons towards the termini were gained together with the ones coding for the gained domain. If there was no significant similarity between these exons and the ones in the sequences without the gained domain, the exons were called novel and the gain still called terminal. Conditions for calling an exon as novel were the following: 85% or more novel amino acids in an exon (i.e. residues

unaligned with amino acids in the sequences without the domain), or less than 10% identity with any of the sequences without the domain. For short exons coding for 20 amino acids or less, the requirement was changed to less than 40% identity. All other domain gains were classified as middle gains.

It is important to note that examining the sequences that surround the gained domains helps to infer the full length of a protein segment that was inserted. In this way, I did not rely solely on domain boundary assignments, which might be imperfect.

3.2.6 Genomic origin of the inserted domain

For all domain gain events that have a human descendant, the gained domain sequence from a representative protein was searched with Wu-blastp against the rest of the human proteome. The best significant hit that was not in one of the gene's paralogues was considered to be a potential donor of the gained domain. A set of paralogs for each gene was composed of other human genes from the same TreeFam family and Ensembl paralogues for that gene. The condition for a significant hit was an E-value of less than 10^{-4} with 60% or more of the domain sequence aligned.

The structures of the genes with gained domains and of their best hits were visually examined using Ensembl (release 50) and the Belvu viewer (<http://sonnhammer.sbc.su.se/Belvu.html>).

The Fisher Exact test in R was used to estimate statistical significance of observed trends (<http://www.r-project.org/>).

The Segmental Duplication Database: <http://humanparalogy.gs.washington.edu/> was used to obtain the coordinates of segmental duplications in the human genome. It was investigated whether any segment from the database spanned any of the representative genes with a domain gain, and if so, whether the other copy of that segmental duplication was placed on the gene that was a potential donor of the domain. It was also checked whether the other copy overlapped with any of the paralogs of the representative gene.

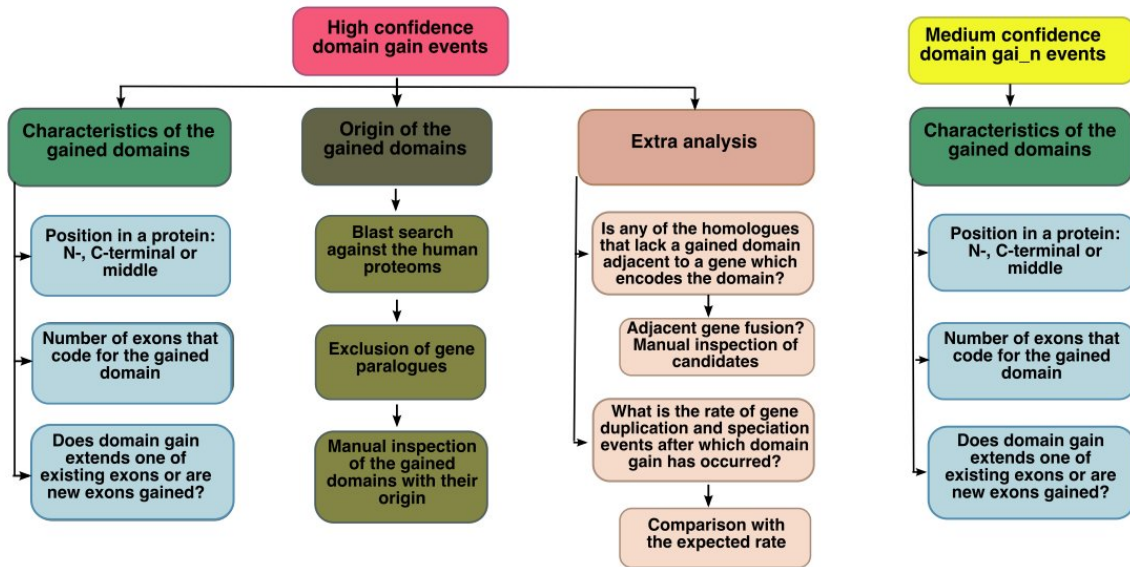


Figure 3.4: Flowchart of analysis for the sets of high and medium confidence domain gain events. For the set of high confidence domain gain events, I looked at characteristics of the gained domains, their potential origin and other trends that could imply potential causal mechanism. For the set of medium confidence domain gain events, I only looked at the characteristics of the domains since this set is enriched with false positives and it was obtained only to test whether the set of high confidence domain gains biased conclusions towards any of the causal mechanisms.

3.3 Results

3.3.1 Set of high confidence domain gain events

To obtain a set of high confidence domain gains I implemented an algorithm that ensured that a gain is not falsely called when other genes in that family had actually experienced multiple losses of the domain in question. I also took into account only those gains that had at least one representative sequence in a genome of better quality and discarded gains where there was only one sequence with the gained domain, i.e. gain was on the leaf of the phylogenetic tree. I did this to overcome the issue of erroneous gene annotations, such as, for example, the instances where two neighbouring genes are annotated as one because regulatory segments that distinguish the genes are not yet identified. Finally, I refined the initial domain assignments to find domains that were missed in the initial Pfam based annotation and discarded all dubious domain gain cases where there was evidence that a domain gain was called due to missing Pfam annotations. After filtering for these confounding factors that could cause false domain gain calls and taking into account only examples where the same transcript contains both the ancestral portion of the gene and a sequence coding for a new domain, I was left with 330 events where I could be confident that one or more domains had been gained by an ancestral protein during animal evolution – I took into account only gains of new domains, and not duplications of existing domains.

The final set is not comprehensive, but these filtering steps were necessary to ensure that the set of domain gain events is of high confidence. Moreover, none of these steps introduces a bias towards any one mechanism over another. The only mechanism of domain gain that I cannot detect after this filtering is the case where amino acid mutations in the sequence created signatures of a domain that was not previously present in the protein; for example, when point mutations in the mammalian lineage created signatures of a mammalian-specific domain.

3.3.2 Characteristics of the high confidence domain gain events

To investigate which molecular mechanisms have caused domain gains in the set of high confidence domain gain events, I examined the characteristics of the sequences that code for the gained domains. As a requirement, each gain event in the set has as descendants two or more genes with the gained domain. To simplify the investigation, I only considered one representative protein for each gain event, and most (232 or 70%) of these were drawn from the human genome as its gene annotation is of the highest quality. Sometimes the same protein was an example for more than one domain gain that occurred during evolution. I projected intron-exon boundaries and intron phases onto the representative protein sequences to help identify the possible causative mechanism. I also compared each representative protein sequence with the orthologs and paralogs in the same TreeFam family that lacked the gained domain. This helped in assigning the characteristics of the gained domains.

I recorded domain gain position (N-, C-terminal or middle) as well as the number of gained exons and whether the domain was an extension of an existing exon (Figure 3.5). I observed two pronounced trends: firstly, most of the domain gains (234 or 71% of the events) occurred at protein termini. This was in agreement with previous studies (Bjorklund et al., 2005; Weiner et al., 2006). Secondly, the majority of the gained domains (again 234 or 71%) are coded for by more than one exon and therefore retroposition is excluded as a likely causative mechanism for them.

I found that different methods for classification of the gain events gave similar results with the most prominent categories of domain gains being gains of multiple novel exons (Appendix B.3). This gave me confidence that domains that are called to be gained on new exons in this analysis indeed are.

Other domains in the same representative proteins that experienced domain gains were also mostly encoded by more than one exon. Namely, 304 out of total 353 domains, or 86% of domains that were present in only one copy in the representative proteins were encoded by two or more exons.

I chose a single representative transcript for each gain event, but as a control, I compared characteristics of the gained domain in all descendant TreeFam transcripts with the domain in the human representative transcript. I

found that in the majority of cases, other descendants of the gain event had the same characteristics of domain gain as the representative protein (on average in 76% descendants of a gain event). This suggests that the causative mechanism can be investigated by looking at the characteristics of the domain in one representative protein for each gain.

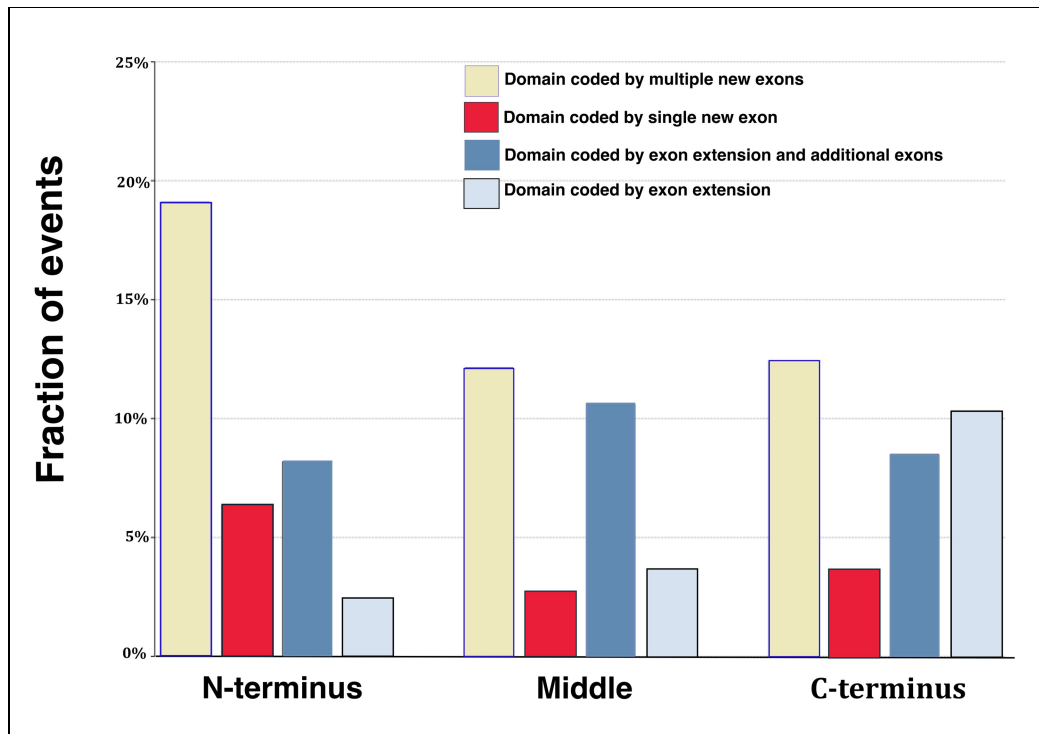


Figure 3.5: Distribution of domain gain events in the high confidence set of domain gains according to the position of domain insertion and number of exons gained. Gains at N- and C- termini and in the middle of proteins are shown separately. The first column in each group shows the fraction of gains where the gained domain is coded by multiple new exons and the second where it is coded by a single new exon. The third column shows the fraction of gains where the ancestral exon has been extended and the gained domain is coded by the extended exon as well as by additional exons. Finally, the fourth column in each group shows cases where only the ancestral exon has been extended with the sequence of a new domain.

3.3.3 Characteristics of the medium confidence domain gain events

The approach for obtaining a set of high confidence domain gains does not bias the final set towards any of the mechanisms. However, the total number of gain events in the set is relatively small and this could introduce apparent dominance of one mechanism over another. Hence, I composed a bigger, but lower confidence, set of events to investigate whether the same trends in domain gains are present in this set; in particular, whether the distribution of characteristics of the gained domains is similar to the one of the high confidence set. I named this set 'Medium confidence' gain events. For this, I used the initially reported set of domain gain events and excluded the filtering criterion which asked for a domain to be present in at least two descendant proteins, and the one which did not allow any similarity between the gained domain and other sequences in the same gene family (Figure 3.3.). I left only the criterion of necessity for domain gains to be supported by a gain in an organism with a better quality genome, since the distribution of domain gains that are reported only in one species – e.g. on the leaf nodes in the trees - showed a bias towards the genomes of lower quality (most gains were reported in *Schistosoma mansoni* and *Tetraodon nigroviridis*: 320 and 303 gains, respectively, and among the organisms with least reported gains were human and mouse: 25 and 19 gains, respectively). I compared the distribution of domains with different characteristics between the high and medium confidence sets of gain events (Figure 3.6). I found that the distribution of domain gains in the two sets is similar overall thus supporting the major conclusions I draw here. The major difference was in the number of middle domains coded by one exon: there were 1.8 times more gains of a domain coded by a single novel middle exon, and 1.6 times more gains of a domain coded by an extension of a middle exon. The set of a medium confidence domain gains is enriched with false domain gain calls caused by discrepancies in the domain annotation of proteins from the same TreeFam families. However, I cannot rule out that a fraction of these gains is real; hence, more supporting cases for the mechanisms that can add domains to the middle of proteins could be found in a larger set. Mechanisms that could be at play here are retroposition and

exonisation of previously non-coding sequence, but also recombination inside the gene sequence.

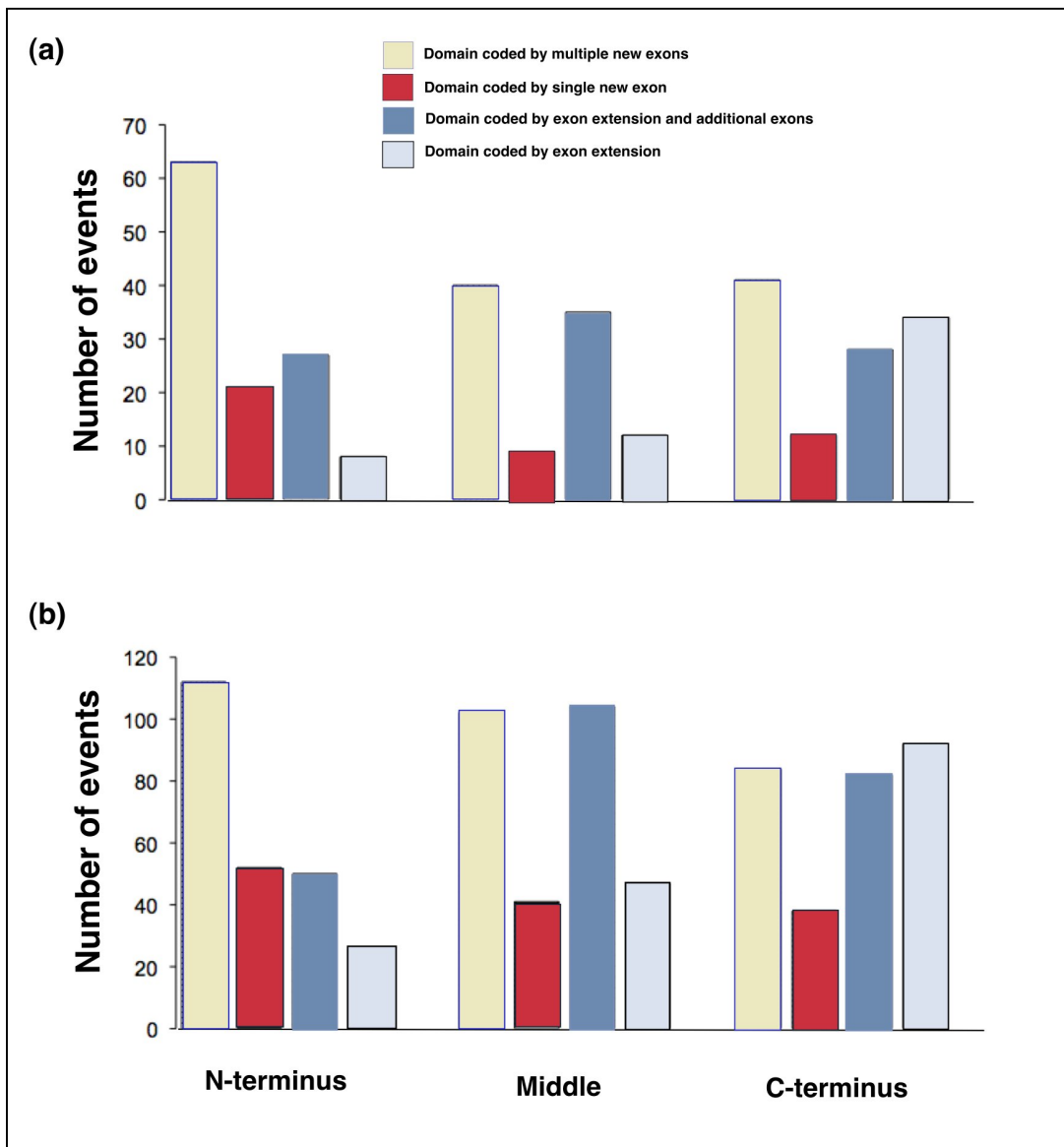


Figure 3.6: Distribution of domain gain events according to the position of domain insertion and number of exons gained in the set of high confidence domain gains and in the set of medium confidence domain gains. Distribution of characteristics of domains from the high confidence set of domain gains (graph a) is for the same – high confidence - gain events represented in Figure 3.5. Graph b) shows the distribution of characteristics of domains from the set of medium confidence domain gains. There are in total 330 high confidence domain gain events and 849 medium confidence domain gains (of which 19 gains have ambiguous position and are not shown in the graph). The flowchart in the Figure 3.3 shows the procedures for creation of these two sets of domain gains.

3.3.4 Supporting evidence for the representative transcripts

I based this work on the Ensembl gene and transcript predictions. However, Ensembl predictions rely on the supporting transcriptome and proteome evidence which is still incomplete. Mistakes in the transcript models can cause false domain gain calls for two reasons: firstly, a transcript that has apparently gained a domain coding sequence can actually exist as two separate transcripts that are falsely annotated as one longer, and secondly, if a domain gain is reported in the genomes with better quality annotations it could be that in the genomes of lower quality the domain is missing only due to incomplete annotation.

To investigate the possible extent of errors introduced by the first type of annotation errors, I checked if there was available supporting evidence for the transcripts that were representatives for domain gain events. I retrieved supporting evidence on the transcript level by using the Ensembl API and checked individual human and mouse representatives without the supporting evidence through the Ensembl website. I found that there was known mRNA supporting the transcript structure in 226 out of 232 human representative gain events and that there were 4 additional cases where evidence was on the exon level. Therefore, 99% (230 of 232) of human representatives have valid supporting evidence. For mouse, there is evidence on the transcript level for 14 out of 18 representative gain cases, and two other transcripts are supported on the exon level. Hence, supporting evidence exists for 89% of the gain events (16 of 18) with mouse representative transcript. For other organisms I took only automatically retrieved transcript evidence into account and I found that in rat there was supporting evidence for 60% (3 of 5) of the events, in chicken and zebrafish for 25% (1 of 4 and 5 of 20 events, respectively), and for frog and fruit fly none of the representative transcripts had available supporting evidence (there were 9 and 43 representative transcripts in frog and fruit fly respectively). It is important to note that the small number of reported gain events with the rat and chicken representative transcripts is possibly also a reflection of the incomplete gene annotations in these species. In conclusion, I

am confident the transcripts with gained domains in human and mouse are correct, but am more cautious about representative transcripts with the gained domain coding sequences in other organisms.

I addressed the level of possible false domain gain calls due to the second type of annotation errors on a smaller set of domain gains which represented a set of gain calls likely to be affected by this error. Namely, domain gains that occurred in the human lineage after the divergence of vertebrates (121 reported domain gain events) can have on one side well studied genomes as human and mouse and on the other side, as an outgroup, lower quality genomes like the one of *C. intestinalis*. For 49 of these gain events the TreeFam family with the reported domain gain also contained orthologous genes in *C. intestinalis* without that domain. I took sequences of *C. intestinalis* orthologs together with 5kb of sequence upstream and downstream of them and performed tBLASTn (<http://blast.wustl.edu/>) to test whether the missing domains were present but only lacked annotation. I found that in four cases at least one of the domains reported to be gained in vertebrates is present in the neighbourhood of *C. intestinalis* orthologous (P-value < 0.1, tBLASTn). However, for two of these cases gene annotation is of very good quality, and the predicted UTR signals and proximity to their neighbouring genes do not support the assumption that the 'missing domains' should be added to these genes. Therefore, I estimate that 4% (2 of 49) of the apparently gained domains could be reported due to errors in gene annotations. However, since these domains are found only in vertebrate genes in the corresponding TreeFam families, these might still be the cases of domain gain but only the time points of the gain events could be before the divergence of *C. intestinalis* from vertebrates. Domains found next to the *C. intestinalis* orthologues, which are possibly missed by incomplete gene annotations were: the Calx-beta domain (PF03160) next to the Ensembl gene ENSCING00000003141 which was gained in the TreeFam family TF105392 together with the Ig-like superfamily (clan CL0159), then the ADP-ribosylation superfamily (clan CL0084) next to the gene ENSCING00000005839 which was gained in the TreeFam family TF329720 together with the BRCA1 C terminus domain (PF00533). The two other domains which were found next to *C. intestinalis* genes with good quality annotation are the Sema domain (PF01403)

next to the gene ENSCING00000006805 - which was gained in TreeFam family TF317402, and the Kunitz/Bovine pancreatic trypsin inhibitor (PF00014) next to the gene ENSCING00000011322 - which was gained in the TreeFam family TF331207.

3.3.5 Donor genes of the gained domains

I investigated whether duplication of the sequence of the 'donor genes' preceded gains of these domains. I selected the 232 gain events with human representative proteins. The selected domain gain events cover those events where at least one of the descendants is a human protein. Hence, the time scale for these events ranges from the divergence of all animals – which was around 700 mya to the divergence of primates – around 25 mya. I grouped descendants of each gain event into the evolutionary group (primates, mammals, vertebrates, bilaterates and animals) they span. In appendix B.2, all gain events together with the information about the evolutionary group of the descendants with the gained domain are listed. I looked for protein regions in the human proteome that are similar to gained domains and, in the case that duplication preceded domain gain could possibly be the source of the gained domains. For this, I used wu-blastp (<http://blast.wustl.edu>). I found a potential origin for 129 (56%) of the gained domains. For the remaining ones it is possible that the mechanism for domain gain either did not involve duplication of an existing 'donor' domain, or that the two sequences have diverged beyond recognition. Hence, the set of domains without the potential 'donor' is enriched in events where the domain has been gained through gene fusion or recombination without previous duplication of the region that encodes the domain or through exonisation of previously non-coding sequence.

3.3.6 Investigation of cellular mechanisms that caused domain gain events

There are several cellular mechanisms, described in the introduction of this chapter, which could have caused the observed domain gain events. I have looked at the characteristics of the gained domains in human representative proteins and attempted to relate these gain events to their possible causative mechanisms.

These gain events illustrate characteristics of domains that were gained during evolution of the human lineage. However, it is important to note that at different stages of evolution different mechanisms could have dominated. The same is valid for domain gains in different species after species divergence. This is why I looked at the characteristics of the gained domains in representative proteins of each species separately. I found that gain of multiple terminal novel exons was a dominant mechanism for domain gains in human, mouse and frog - these gains made 34, 50 and 56%, respectively of all gains with representative protein in these species. In fruit fly, the dominant category of gains was extension of exons at C-terminus - 29% of domain gains - and dominant gains in zebrafish were a mixture of two - 35% of gains were novel terminal domains and 20% C-terminus exon extensions. For rat and chicken there were too few domain gains for me to draw conclusions.

3.3.6.1 Retroposition as a mechanism of domain gain

Domains in the human lineage for which I could identify a potential donor protein and which are gained within a single exon are possible candidates for retroposition (26 cases). I further investigated these gain events. Retroposition would be supported as a causative mechanism if there were no other exons gained together with the one that encodes the new domain, and also if a long interspersed nuclear element (LINE) retrotransposon was present before the gained domain and/or 'donor' domain. Inspection of the candidate domains showed the supporting evidence for the gain of pre-SET and SET domains in the

SETMAR gene by this mechanism (described in Figure 3.7) but not for other candidate gained domains. However, this inspection was hampered with the fact that the gained domain often existed in multiple copies in the 'donor' protein so it was difficult to judge which of the domain repeats was the potential origin. Finally, in the cases where extra exons appeared to be gained with the one that encodes the new domain, retroposition could be excluded as a likely mechanism. The lack of a LINE element does not rule out retroposition as a possible mechanism, rather it does not show additional support for it. Even if isolated, the example of the SETMAR gene is very relevant, since there are only a few cases reported of the role of retroposition in the creation of novel genes in the human lineage (Babushok et al., 2007a). The pre-SET and SET domains in the SETMAR gene most likely have an origin in the gene SUV39H1. Interestingly, the SETMAR gene lies in the intron of another gene (SUMF1) and hence possibly uses its regulatory mechanism for transcription.

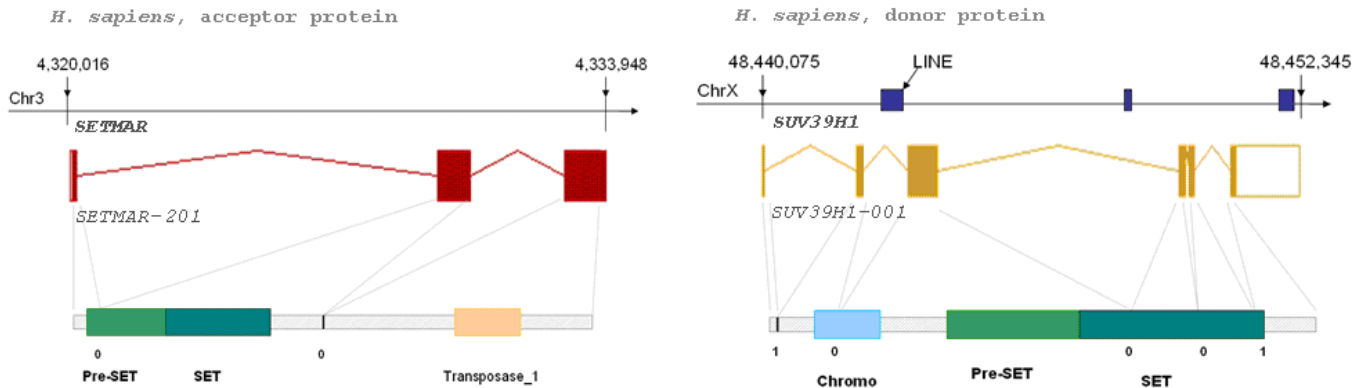


Figure 3.7: Retroposition as a causing mechanism for domain gain.

An example of a domain gain mediated by retroposition. TreeFam family TF352220 contains genes with a transposase domain (PF01359). The primate transcripts in this family have been extended at their N-terminus with the pre-SET and SET domains. The representative transcript for this gain event is SETMAR-201 (ENST00000307483, left in the figure). Both gained domains have a significant hit in the gene SUV39H1 (ENSG00000101945, right in the figure - the Set domains of the donor and recipient proteins share 41% identity). Previously, it has been reported that the chimeric gene has originated in primates by insertion of the transposase domain (PF01359, with a mutated active site and no transposase activity) in the gene that had had the pre-SET and SET domains (Cordaux et al., 2006). Here, I propose that the evolution of this gene involved two crucial steps: retroposition of the sequence coding for the pre-SET and SET domains and insertion of the MAR transposase region described by Cordaux et al. The SET domain has lost the introns present in the original sequence and the Pre-SET domain has an intron containing repeat elements in a position not present in the original domain suggesting it was inserted later on. The likely evolutionary scenario here includes duplication of pre-SET and SET domains through retroposition, insertion of transposase domain and subsequent joining of these domains. The SETMAR gene is in the intron of another gene (SUMF1), which is on the opposite strand so it might be that SETMAR is using the other gene's regulatory regions for its transcription. The top of the figure shows the genomic position of depicted genes. Arrowheads on the lines that represent chromosomal sequences indicate whether the transcripts are coded by the forward or reverse strand. Transcripts are always shown in the 5' to 3' orientation and proteins in the N- to C-terminal orientation. Exon projections and intron phases are also shown on the protein level. Pfam domains are illustrated as coloured boxes. Figures 3.8 and 3.9 use the same conventions.

3.3.6.2 Joining of adjacent genes as a mechanism of domain gain

Terminal gains of domains coded by multiple novel exons are particularly interesting because for these events there is only one plausible causative mechanism: joining of exons from adjacent genes (Figure 3.2). Because of the criteria I used here, the number of new exons gained at termini is a lower estimate. Nonetheless, this is still the most abundant type of event. 104 or 32% of all events are N-terminal (63 events) or C-terminal (41 event) gains of domains coded by multiple new exons (Figure 3.5). I can discard retroposition and recombination assisted insertions into introns as likely mechanisms for these gains. However, it is possible that recombination preceded domain gains, and even that recombination did not juxtapose fully functional genes but only, for example, certain exons of one or both of the genes. Indeed, I have not found that these genes exist as adjacent separate genes in the modern genomes (described below) and it is likely that these gains were preceded by DNA recombination.

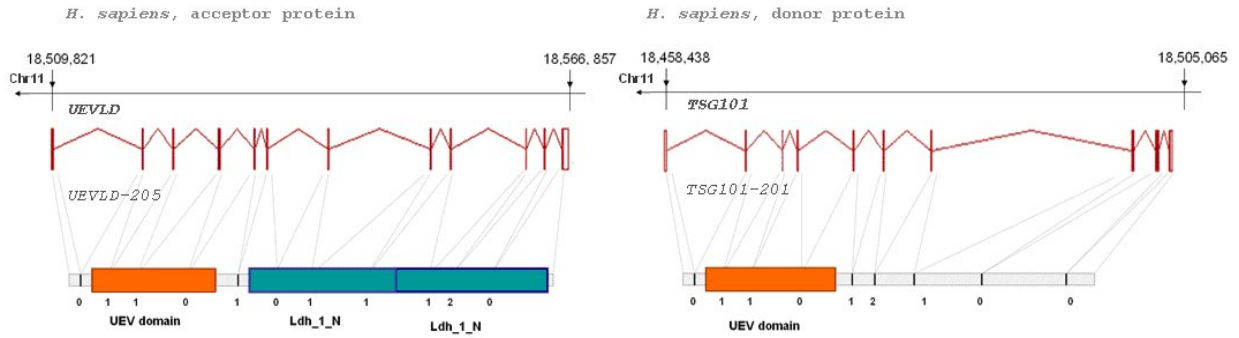
The search for the 'donor gene' of the gained domains identified the possible origin of the domain for 60% of domains coded by new terminal exons. This implies that duplication of a donor domain has frequently provided the material for subsequent exon joining and new exon combinations. An illustration of this mechanism is the gain of the UEV domain in the UEVLD gene (Figure 3.8 and 3.9). The gain has most likely occurred after the neighboring gene TSG101 has been duplicated and exons of one copy joined with the UEVLD ancestor's exons. Two similar examples, for the evolution of genes CELSR3 and AC093283.3, are also illustrated in Figure 3.8.

Gains of multiple novel terminal exons make up 32% of all domain gains and are best explained with joining of adjacent exons. On the other hand, terminal gains of domains coded by a single novel exon can be explained either by the joining of exons from adjacent genes or with other mechanisms such as retroposition. The former mechanism is more likely since, together with the novel exon that codes for the gained domain, extra exons, that do not code for the gained domain, have frequently been gained (in at least 42% events, or 18 of total 42 cases). Also, further inspection of the candidate gains in the human

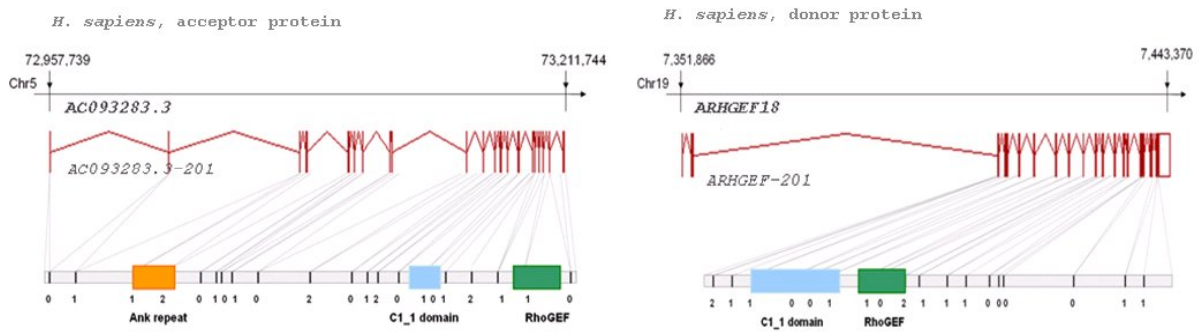
lineage did not find LINE elements that preceded a gained or 'donor' domain and hence did not lend support for retroposition as a causative mechanism (described above). With regard to other categories of domain gain events in Figure 3.5., because of the strict criteria I used to call a gained domain terminal and coded by novel exons, a number of exon extensions and middle gains are possibly misclassified terminal gains and gains of novel exons.

Recent segmental duplications in the human genome are a possible source of new genetic material (Bailey et al., 2002) and their role in the evolution of primate and human specific traits has been debated (Bailey and Eichler, 2006). Hence, I investigated whether recent domain gains in the human lineage could be related to the reported segmental duplications. I found two domain gains that were best explained by recent segmental duplications and subsequent joining of two genes (Figure 3.10). Both of these gains occurred at the protein termini after divergence of primates. The mechanism of their evolution is the same as in the case of the UEVLD gene: joining of exons from adjacent genes after gene duplication. Additionally, for these two examples, there is also evidence of a likely connection between recent genomic duplication and domain gain. In spite of this, it is necessary to be cautious when assessing the possible role of these proteins. For both examples, there is only transcript evidence and some of the transcript products of these genes appear to have a structure that would lead to them being targeted by nonsense mediated decay (NMD) (Wilming et al., 2008). However, it is still not sure if these genes are targets for NMD or not.

(a)



(b)



(c)

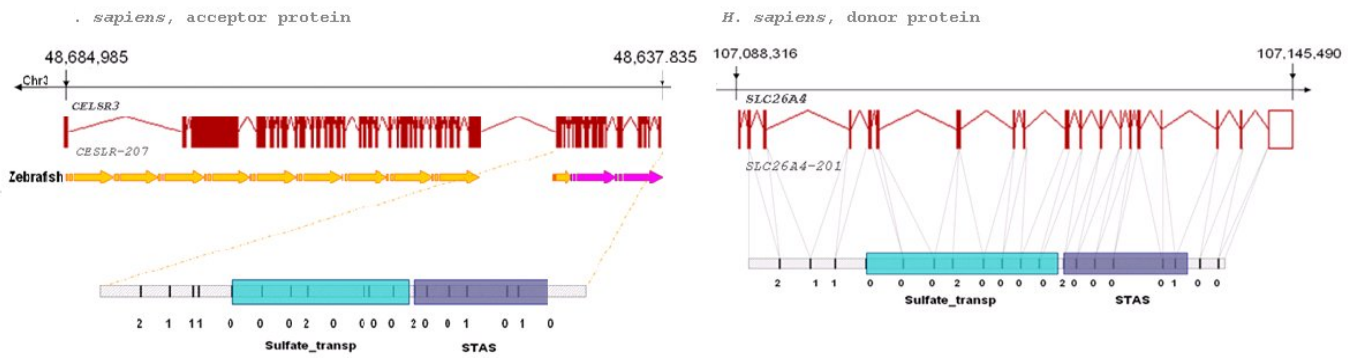


Figure 3.8: Examples for domain gains by joining of exons from two ancestral genes. A representative protein for a domain gain is always shown on the left and a protein which is a potential origin of the gained domain is shown on the right. (a) An example of a domain gain by gene duplication followed by exon joining. TreeFam family TF314963 contains genes with lactate/malate dehydrogenase domain where one branch with vertebrate genes has gained the additional UEV domain. Homologues, both orthologues and paralogues, without the gained domains are present in a number of animal genomes. A representative transcript with the gained domain is UEVLD-205 (ENST00000396197, left in the figure). The UEV domain in that transcript is 56% identical to the UEV domain in the transcript TSG101-201 (ENST00000251968) that belongs to the neighboring gene TSG101 and the two transcripts also have introns with identical phases in the same positions. The likely scenario is that after the gene coding for the TSG101-201 transcript was duplicated, its exons have been joined with the ones of the UEVLD-205's ancestor and the two genes have been fused.

(b) Another example for a domain gain after gene duplication and exon joining. Family TF334740 in the TreeFam database contains genes that code for the Rho-guanine nucleotide exchange factor (RhoGEF). However, the RhoGEF domain was not present in the ancestral protein but was inserted later on together with the C1_1 domain when mammals diverged from other vertebrates (TreeFam release 6.0 that we used in the analysis had chicken, fish and frog genes without the gained domains). The representative transcript for the gain event is AC093283.3-201 (ENST00000296794). The gene ARHGEF18 (ENSG00000104880) has both of these domains, and the two RhoGEF domains between the genes are 52% identical. Hence, ARHGEF18 is a plausible donor for this gain event. Again, the mechanism for the gain of these domains most likely involves gene duplication and exon joining.

(c) TreeFam family TF323983 contains 'Cadherin EGF LAG seven-pass G-type receptor (CESLR) precursor genes. One branch of the family, containing vertebrate genes, has gained the Sulfate transport and STAS domains in addition to the ancestral cadherin, EGF and other extracellular domains. The gain occurred after the other vertebrates diverged from fish, and homologues without the gained domains are present in all animals. A representative for the gain is the transcript CELSR3-207 (ENST00000383733) and its 3' end is shown left in the figure (the whole transcript is too long to be clearly presented). Right in the figure is shown a gene that is the plausible donor of these domains. Namely, the gene SLC26A4 (ENSG00000091137) contains both domains, and its STAS domain is 31% identical to the one in the CELSR3 gene. In addition, the alignment with the Zebrafish genome is shown below the CELSR3-207 transcript. The yellow arrows represent the alignment with the chromosome 8 in Zebrafish, and pink arrows with the chromosome 6 (information taken from the USCS browser: <http://genome.ucsc.edu>). The alignment with the fish genome shows that the synteny is broken exactly in the region where the new domain is gained. Therefore, the plausible scenario for domain gain involves gene duplication, recombination and joining of newly adjacent exons.

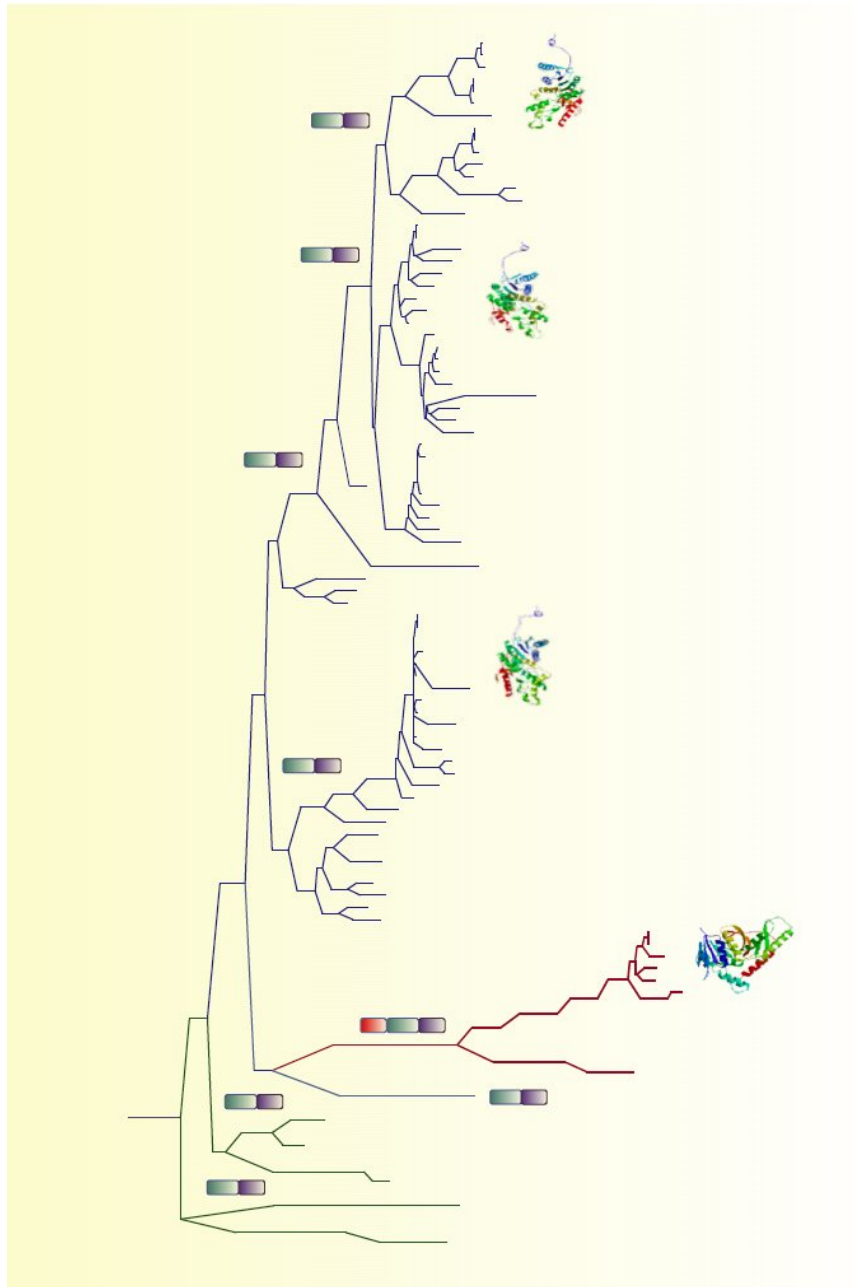
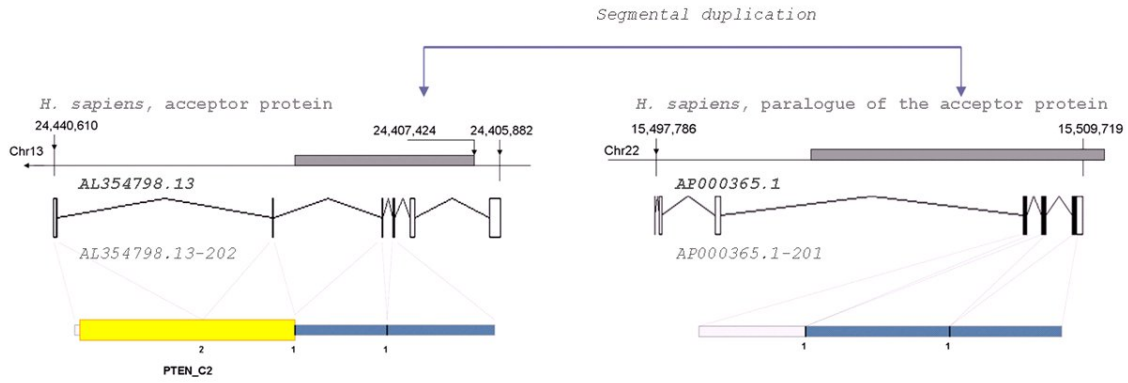


Figure 3.9: Gain of the UEV domain in the TreeFam family TF334740. Structure of a representative gene that was extended with the UEV domain is shown in Figure 3.8a. Here, the evolutionary tree of lactate dehydrogenase genes is shown. Vertebrate genes in the tree – the red coloured branch – have gained the UEV domain during evolution. This should influence both protein structure and function. Models of the protein structures of example proteins in different branches of the tree are shown. The structure is predicted from protein sequence, based on similarity with proteins with solved structures, using Swiss-model (<http://swissmodel.expasy.org>). The domain gain occurred after gene duplication and subsequent joining of exons from adjacent genes, which appears to be the dominant mechanism for acquiring new domains during animal evolution.

(a)



(b)

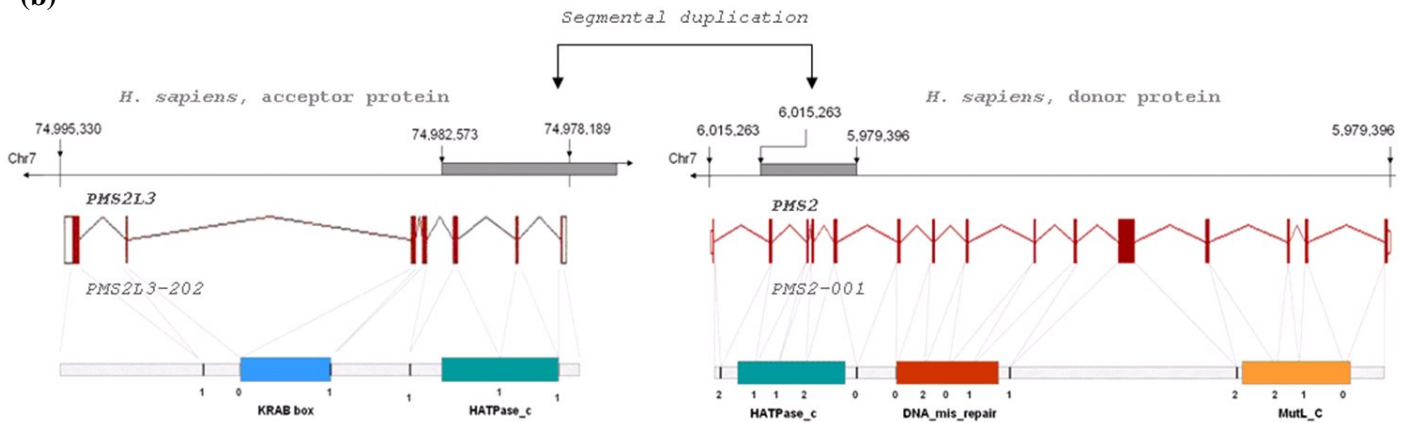


Figure 3.10: Examples for domain gains by joining of exons from adjacent genes assisted by recent segmental duplication. (a) An example for a domain gain after segmental duplication and exon joining. TreeFam family TF351422 contains only primate genes, and after a gene duplication event one branch of the family has gained the PTEN_C2 domain. A representative transcript for this gain is AL354798.13-202 (ENST00000381866). There are few segmental duplications spanning across the gene AL354798.13 and one of them is covering only the ancestral portion of the gene – without the gained domain. The pair of that segmental duplication is on the gene's paralog that has not gained the domain, the gene AP000365.1 (ENSG00000206249). Hence, a possible scenario is that a recent duplication of a paralog gene has changed its genetic environment and brought it to the proximity of the PTEN_C2 domain which subsequently became part of the gene.

(b) Another example of a gain of a domain coding region by segmental duplication followed by exon joining. A branch with primate genes in the TF340491 family of vertebrate proteins that contains the KRAB domain has gained the additional HATPase_c domain. The representative transcript is the human PMS2L3-202 (ENST00000275580). The HATPase_c domain exists in the gene PMS2 (ENSG00000122512) and on the protein level the gained domain is 98% identical to the sequence in the protein product of the PMS2's transcript PMS2-001. There is a segmental duplication that spans across the gained sequence in the transcript PMS2L3-202 and is a pair of the segmental duplication that covers the same domain in the gene PMS2. The pair of segmental duplication regions are presented as grey boxes and connected with arrows. Therefore, the mechanism underlying this gain appears to be a segmental duplication of the sequence belonging to PMS2 after which the copy next to the PMS2L3-202's ancestor was joined with it. An important caveat is that PMS2L3-202 has a structure that can be targeted by NMD.

3.3.6.3 Insertion of exons into ancestral introns as a mechanism of domain gain

Because of the special attention that has been given to domain insertions into introns in discussions on exon shuffling (Liu and Grigoriev, 2004; Patthy, 1999), I have studied the middle gains of novel exons in more detail. The theory of domain shuffling by intronic recombination states that the exons inserted into ancestral introns are surrounded by introns of symmetrical phases (Patthy, 1999). I looked at the phases of introns surrounding the domains inserted into the ancestral introns. A list of all intronic gains is in Appendix B.4. Twenty six of them had the agreeing phases on the boundaries of exons that encoded them, and two more were gained with extra exons that also had agreeing phases on boundaries. Only one in three possible intron phase combinations gives the same intron phases, and here I observed a strong bias in agreement of intron phases surrounding the gained domains (57% or 28 out of 49 domains are surrounded with introns of the same phase) and among these I also observed an excess of 1-1 phases on exon borders (79% or 22 out of 28). Both symmetrical phases and an excess of 1-1 phases are considered to be supporting evidence for intronic insertions (Patthy, 1999). Moreover, intronic insertions have been shown to be widespread in extracellular matrix proteins and the gained domains in this subset of domains are well known extracellular domains (such as EGF, Sushi, Fibronectin and Immunoglobulin domains) (Patthy, 1999). However, these potential examples for domain insertions into introns cover less than 10% of all gain events; which does not support the expectation that this was the major mechanism for domain gains in the evolution of metazoa (Kaessmann et al., 2002; Liu and Grigoriev, 2004). It is also worth noting that the majority (82% or 40 of 49 intronic gains) of domains inserted into ancestral introns were coded by multiple exons, which implies that intronic recombination, rather than retroposition, would be more likely the causative mechanism for the majority of intronic gains. In conclusion, the majority - 28 out of 49 - domains coded by novel exons and gained into the middle of proteins are surrounded by introns of symmetrical phases, and hence give support to the assumption that the causative mechanism for them included insertions into ancestral introns.

Related to exons insertions into introns; it has been shown that a class of domains whose borders strongly correlate with their encoding exon borders had experienced significant expansion during animal protein evolution (Liu et al., 2005). Moreover, these domains were also found to be frequent in novel metazoan multidomain architectures (Ekman et al., 2007). It has been hypothesised that these domains have contributed to exon shuffling in metazoa (Liu et al., 2005) and a correlation with symmetrical intron phases surrounding these domains was attributed to their intronic insertions (Liu et al., 2005). I investigated how well represented these domains were in the set of high confidence domain gain events. I found that they make up about 28% of the set (101 out of 362 gained domains, or 97 out of 333 gain events) which is a significant overrepresentation since only 103 out of total 8,634 domains or clans in the Pfam 23 are in the class of exon-bordering domains (1.2% of all domains). The significant fraction of these domains in the dataset confirms their important role in domain shuffling in metazoa, but the fact that they have been gained about as equally frequently at N- or C-terminus as in the middle of proteins (35, 30 and 32 events, respectively) shows that they have been important not only for intronic gains, but for domain rearrangements in animals in general.

3.3.6.4 Exonisation of previously non-coding sequences as a mechanism of domain gain

Figure 3.5. shows that a relatively high fraction of domain gains occurred as extensions of C-terminus exons. If exonisation of a previously non-coding sequence was a causal mechanism for some of the domain gains, one would expect that these gains would preferentially occur as extension of exons at C-termini. Extensions of exons at N-termini and in the middle of proteins have a risk of introducing a frame-shift and being selected against. Additionally, one would expect that when a new Pfam family is formed from previously non-coding sequence (by exon extension) that it is more likely that this will be an intrinsically unstructured region. Intrinsically unstructured or disordered regions do not have a stable globular structure, but are associated with important functions (Wright and Dyson, 1999; Gsponer and Babu, 2009;

Gsponer et al., 2008). I predicted disordered regions in all proteins from the study with the IUPred software (Dosztanyi et al., 2005) and looked at the average percentage of disordered residues in each gained domain in the set (Figure 3.11) and in all other domains present in these proteins. I observed two prominent trends: firstly, gained domains in general have a greater percentage of disordered residues (on average only 5% of residues of all other domains in proteins are predicted to be disordered compared to on average 21% of residues in the gained domains) and secondly, domains with the greatest percentage of disordered residues are those that have been gained by extension of existing exons.

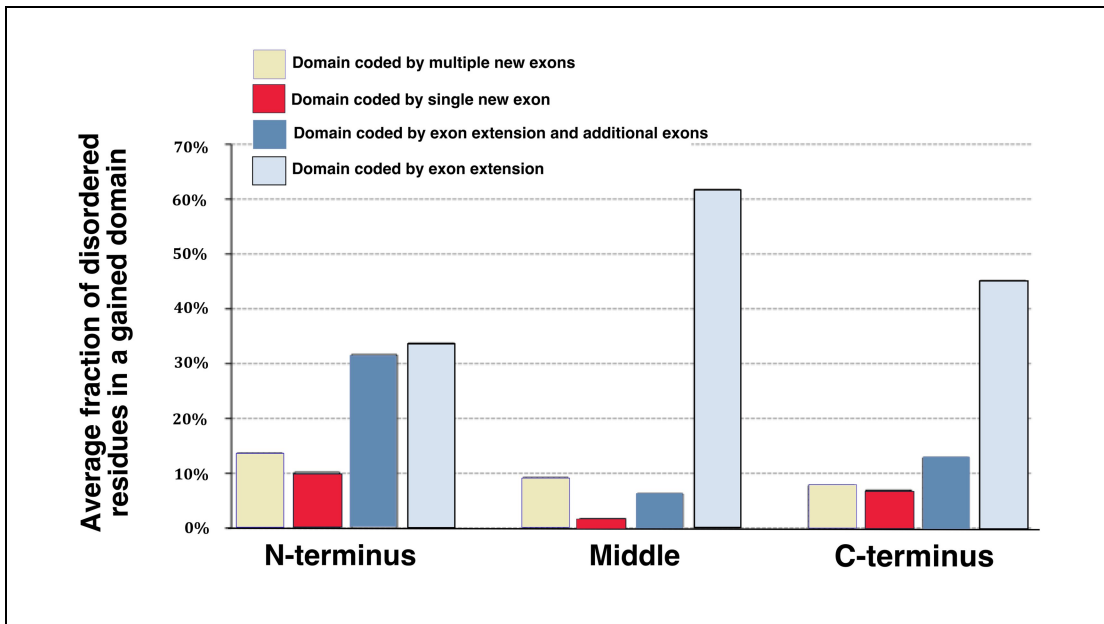


Figure 3.11: Distribution of disordered residues in the gained domains according to the position of domain insertion and number of exons gained. This graph shows the percentage of disordered residues in each category of domain gains. The number of events in each category can be seen in Figure 3.5.

Next, I investigated the individual examples for domain gains through extension of C-terminal exons in the human lineage. By looking at the alignments for these gains, it was possible to find four convincing events of true exon extensions. None of these had a potential 'donor gene' identified in the human proteome. Further inspection of these domains showed that they have actually occurred at that point in the evolution for the first time and the possible mechanism for inclusion of these novel domains was reading through the stop signal and exonisation of previously non-coding sequences (for the gains in primates and mammals alignments at the UCSC genome browser (Kent et al., 2002) show similarity of the gained domains with non-coding regions in the genomes of non-primates and non-mammals, respectively). These examples are: (1) Gain of a proline rich Pfam family PF04680 in primates – in the TreeFam family TF331377, (2) gain of a selenoprotein P C-terminal Pfam family PF04593 in mammals – in the TreeFam family TF333425, and gain of the families: (3) connexin 50 C-terminal - PF03509 and (4) the Kv2 voltage gated K⁺ channel - PF03521 in vertebrates – in the TreeFam families TF329606 and TF313103, respectively. Representative transcripts for these gains can be found in Appendix B.2. It is noteworthy that none of these Pfam families has a solved structure and it is possible that they are not true structurally independent protein domains. Even so, their sequences are conserved in the organisms in which these Pfam families are present (it was possible to recognize these domains in the sequence), which implies that they could be functionally relevant.

3.3.7 Domain gains most frequently occur after gene duplications

One advantage of using TreeFam phylogenies is the ability to distinguish between gene evolution that follows gene duplication and the one that follows speciation. I investigated whether there was any correlation between domain acquisition and gene duplication. In the entire database, speciation nodes are more frequent than duplication nodes (there are 3.43 times more internal speciation nodes; in total there are 394,853 internal speciation and 115,013 internal duplication nodes). However, in the set of domain gain events that have

a human representative for the gain, duplication nodes were more frequent (a change in domain architecture was 1.32 times more frequent after gene duplication; 101 gain events occurred after speciation event and 133 after gene duplication). Hence, when comparing the observed versus expected frequency of domain gains after duplication and speciation events I found that domain gains occurred nearly five times more frequently than expected (1.32 relative to 0.29). As a control, I also checked the branch lengths after speciation and duplication nodes and found that domain gains occurred after every 3,455 units of branch length when the event was speciation and after 1,274 units of length when the event was duplication. Hence, the lower estimate is that domain gains occurred 2.72 (~3) times more frequently after gene duplication compared to after speciation. This shows that not only duplication of the 'donor gene', but also of the 'recipient gene' assisted domain gains. Taken together with the gain events that had the 'donor genes' identified, in 80% of the domain gains, duplication of either the ancestral protein or donor protein has been involved. Moreover, when two genes were fused together then the assignment of 'donor' and 'recipient' genes depends solely on whose phylogeny is one looking at.

When I grouped the gain events with the identified 'donor genes' according to the age of the event and looked at the chromosomal position of the 'donor genes' I observed a trend that in the human lineage the younger the gain event was, the more likely it was that the 'donor gene' would be found on the same chromosome (Figure 3.12). However, the numbers of domains found on the same chromosomes are small (Figure 3.12). Therefore, I grouped values for domain gains before and after divergence of mammals and found that in spite of the small set of domain gains, the difference in trend is still present (P-value = 0.03, Fisher exact test). The fact that the tendency was decreasing for the older gains could be related to continuous chromosomal rearrangements. In addition to that, I observed that in general the 'donor genes' were found on the same chromosomes as the genes with the gained domains more frequently than would be expected by chance. I calculated this as follows: I compared the number of gains on each chromosome with the number of best hits that I would expect to observe if the duplicates could be inserted equally likely anywhere in the genome (calculated as the portion of the genome length on each chromosome –

i.e. individual chromosome length divided by the total length of all autosomes together with X and Y chromosomes - times number of gains on that chromosome). The number of observed 'donor genes' on the same chromosome, 16, is 2.5 times higher than the expected 6.5. This suggests that the duplication mechanism favored creation of duplicates on the same chromosomes.

However, not all domain gains rely on gene duplication. As already discussed, exonisation of previously non-coding sequence does not have to be preceded by gene duplication. Additionally, a closer look at domain gains after primate divergence showed that two domain gain events are actually gains of transposon (CL0219 in the TF328297 TreeFam family) and retroviral (CL0074 in the TF331083 TreeFam family) domains. Gains of domains from mobile genetic elements can also be relevant for the evolution of protein function (Cordaux et al., 2006) and are not necessarily connected with gene duplication.

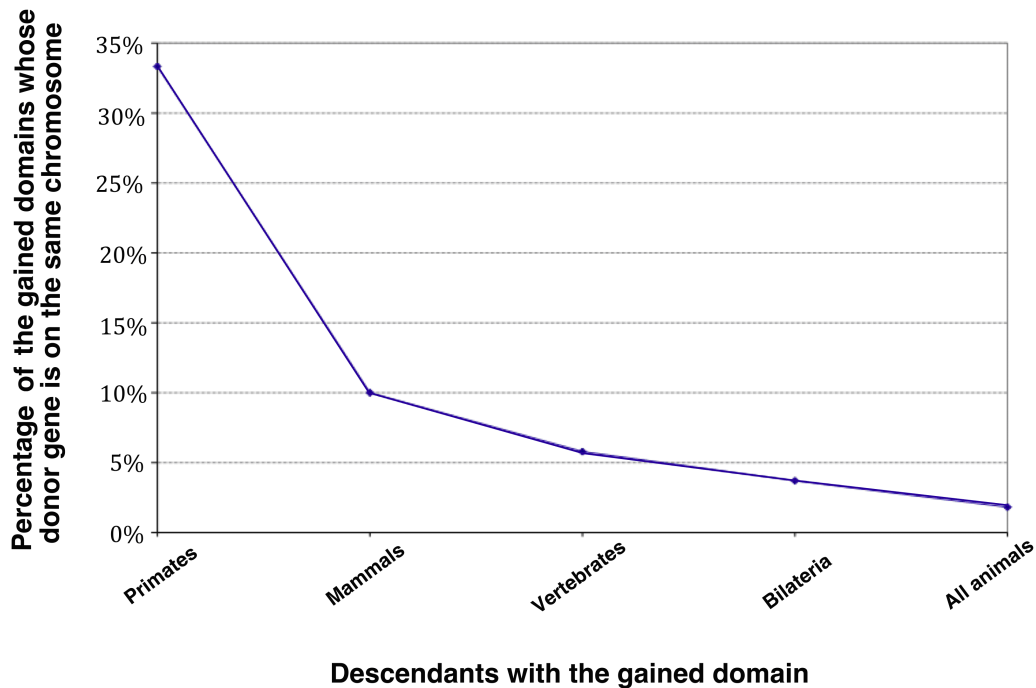


Figure 3.12: Chromosomal position of the ‘donor gene’ and the relative age of the gain event. The graph is showing the fraction of events for which the ‘donor gene’ of the gained domain is identified, and is on the same chromosome as the gene with the gained domain, with respect to the relative age of the gain event. The gain events were divided into five groups according to the expected age of the event as judged by the TreeFam phylogeny. The X axis shows the evolutionary group in the human lineage which descendants of the gain event belong to, and the Y axis percentage of gain events in each evolutionary group for which both of the conditions were valid: I was able to find the donor gene and the donor gene was on the same chromosome as the gene with the gained domain (3 out of 9 gain events in Primates, 2 out of 20 in Mammals, 7 out of 121 in Vertebrates, 1 out of 27 in Bilateralialia and 1 out of 55 gain events in all animals). Appendix B.2 has information about domain gain events that belong to each phylogenetic group. Estimated divergence times (in million years ago – mya, as taken from Ponting (Ponting, 2008) are the following: 25 mya for Primates, 166 for Mammals, 416 for Vertebrates and 700 for all animals (we were not able to estimate divergence time for Coelomata).

3.3.8 Gained domains do not have their origin in the adjacent genes

When a domain gain occurred through joining of exons from adjacent genes then it is possible that this process was assisted with gene recombination, which juxtaposed the sequences of the two ancestral genes together. Alternatively, it is possible that the 'donor' gene with the gained domain was adjacent to the 'acceptor' gene for a long period of time and then in a certain evolutionary lineage the two genes fused. I investigated whether there were instances where a homologue, which lacked the domain, had a gene coding for the gained domain adjacent to it. I found three cases in the present animal genomes where a homologue of a gene with a gained domain did not have that domain but was annotated adjacent to the gene which encoded the domain. If these were true separate genes, these would be examples for joining of exons from adjacent genes and subsequent gene fusion. However, further inspection showed that they were most likely results of gene annotation discrepancies and were possibly not even true domain gains. Therefore, I excluded these gain events from the set of high confidence domain gains. These were the following gains: gain of the BRCA1 C Terminus domain (PF00533) in the TreeFam family TF329705, gain of Kuntiz/Bovine pancreatic trypsin inhibitor (PF00014) in the TreeFam family TF316148 and gain of the LEM domain (PF03020) in the TreeFam family TF317729. In conclusion, for the obtained set of gain events, there is no evidence in the current animal genomes that the gained domains had an origin in the genes that were for long evolutionary times adjacent to the ancestors without the gained domains.

3.3.9 Domain gain events affect cellular regulatory networks

It has been proposed that the novel combinations of preexisting domains had a major role in the evolution of protein networks and more complex cellular activities (Pawson and Nash, 2003; Peisajovich et al., 2010). In agreement with this, I found that the most frequently gained protein domains in the human lineage - domains independently gained 5 or more times in the set of confident

gain events - are all involved in signaling or regulatory functions; the Ankyrin repeat (gained 6 times) and SAM domain (gained 5 times) are commonly involved in protein-protein interactions, and the Src homology-3 and PH domain-like superfamily (both gained 6 times) have frequently a role in signaling pathways. Furthermore, I used the DAVID service (Dennis et al., 2003) to investigate if human representative transcripts (from the table in Appendix B.2) were enriched in any GO terms. Significantly enriched GO terms are listed in Table 3.1, and are in general involved in signal transduction; among the significant terms are 'adherens junction', 'protein modification process' and 'regulation of signal transduction'. This further supported the role of novel domain combinations in the evolution of more complex regulatory functions.

Table 3.1: Significant GO terms (P-value < 0.05 after correcting for multiple testing) for human genes that have been extended with a new protein domain. GO terms are obtained and clustered by using the DAVID service. Abbreviation CC is for Cellular Component, BP for Biological Process and MF for Molecular Function. EASE P-values represent modified Fisher exact P-values. 'Benjamini' shows P-values after applying the Benjamini correction for multiple tests.

	Category	GO term ID	GO term description	EASE P-Value	Benjamini
Annotation Cluster 1	CC	0016323	basolateral plasma membrane	1.1 x10 ⁻⁶	3.1 x10 ⁻⁴
	CC	0005924	cell-substrate adherens junction	4.3 x10 ⁻⁵	5.8 x10 ⁻³
	CC	0030055	cell-substrate junction	6.3 x10 ⁻⁵	5.8 x10 ⁻³
	CC	0005925	focal adhesion	2.3 x10 ⁻⁴	1.3 x10 ⁻²
	CC	0005912	adherens junction	5.9 x10 ⁻⁴	2.7 x10 ⁻²
	CC	0070161	anchoring junction	1.2 x10 ⁻³	4.5 x10 ⁻²
Annotation Cluster 2	BP	0006793	phosphorus metabolic process	5.4 x10 ⁻⁶	9.2x10 ⁻³
	BP	0006796	phosphate metabolic process	5.4 x10 ⁻⁶	9.2x10 ⁻³
	MF	0030554	adenyl nucleotide binding	5.6 x10 ⁻⁶	8.4 x10 ⁻⁴
	BP	0043687	post-translational protein modification	6.2 x10 ⁻⁶	5.3 x10 ⁻³
	MF	0001883	purine nucleoside binding	8.2 x10 ⁻⁶	7.4 x10 ⁻⁴
	MF	0001882	nucleoside binding	9.7 x10 ⁻⁶	7.3 x10 ⁻⁴
	MF	0005524	ATP binding	1.5 x10 ⁻⁵	9.6 x10 ⁻⁴
	MF	0032559	adenyl ribonucleotide binding	2.1 x10 ⁻⁵	1.2 x10 ⁻³
	MF	0003824	catalytic activity	7.5 x10 ⁻⁵	3.1 x10 ⁻³
	BP	0006468	protein amino acid phosphorylation	8.6 x10 ⁻⁵	3.6 x10 ⁻²
	BP	0043412	biopolymer modification	1.1 x10 ⁻⁴	3.8 x10 ⁻²
	BP	0019538	protein metabolic process	1.4 x10 ⁻⁴	3.4 x10 ⁻²
	BP	0006464	protein modification process	2.0 x10 ⁻⁴	3.7 x10 ⁻²
	MF	0017076	purine nucleotide binding	2.7 x10 ⁻⁴	8.2 x10 ⁻³
	MF	0004672	protein kinase activity	5.9 x10 ⁻⁴	1.4 x10 ⁻²
	MF	0032553	ribonucleotide binding	8.0 x10 ⁻⁴	1.7 x10 ⁻²
	MF	0032555	purine ribonucleotide binding	8.0 x10 ⁻⁴	1.7 x10 ⁻²
	MF	0004713	protein tyrosine kinase activity	1.9 x10 ⁻³	3.5 x10 ⁻²
	MF	0016301	kinase activity	2.1 x10 ⁻³	3.7 x10 ⁻²
	MF	0000166	nucleotide binding	2.2 x10 ⁻³	3.6 x10 ⁻²
MF	0016772	transferase activity, transferring phosphorus-containing groups	2.8 x10 ⁻³	4.0 x10 ⁻²	
Annotation Cluster 3	MF	0008270	zinc ion binding	7.3 x10 ⁻⁴	1.6 x10 ⁻²
	MF	0043169	cation binding	1.9 x10 ⁻³	3.6 x10 ⁻²
	MF	0046872	metal ion binding	2.3 x10 ⁻³	3.6 x10 ⁻²
	MF	0043167	ion binding	2.8 x10 ⁻³	4.2 x10 ⁻²
	MF	0046914	transition metal ion binding	2.9 x10 ⁻³	4.0 x10 ⁻²

Annotation Cluster 4	MF	0005088	Ras guanyl-nucleotide exchange factor activity	2.9 x10 ⁻⁶	6.5 x10 ⁻⁴
	MF	0005089	Rho guanyl-nucleotide exchange factor activity	6.9 x10 ⁻⁶	7.7 x10 ⁻⁴
	BP	0035023	regulation of Rho protein signal transduction	5.4 x10 ⁻⁵	3.0 x10 ⁻²
	MF	0005085	guanyl-nucleotide exchange factor activity	2.3 x10 ⁻⁴	7.2 x10 ⁻³
	MF	0030695	GTPase regulator activity	4.1 x10 ⁻⁴	1.1 x10 ⁻²
	MF	0060589	nucleoside-triphosphatase regulator activity	5.1 x10 ⁻⁴	1.3 x10 ⁻²
	MF	0005083	small GTPase regulator activity	1.3 x10 ⁻³	2.6 x10 ⁻²
	MF	0030234	enzyme regulator activity	2.3 x10 ⁻³	3.5 x10 ⁻²
Annotation Cluster 5	MF	0046030	inositol trisphosphate phosphatase activity	1.9 x10 ⁻⁴	6.7 x10 ⁻³
	MF	0004445	inositol-polyphosphate 5-phosphatase activity	1.9 x10 ⁻⁴	6.7 x10 ⁻³
Annotation Cluster 6	MF	0004386	helicase activity	1.2 x10 ⁻⁴	4.5 x10 ⁻³
	MF	0070035	purine NTP-dependent helicase activity	2.1 x10 ⁻³	3.6 x10 ⁻²
	MF	0008026	ATP-dependent helicase activity	2.1 x10 ⁻³	3.6 x10 ⁻²
Other significant GO terms	MF	0005044	scavenger receptor activity	2.6 x10 ⁻⁶	1.2 x10 ⁻³
	MF	0019992	diacylglycerol binding	3.6 x10 ⁻⁵	1.8 x10 ⁻³
	MF	0005488	binding	6.7 x10 ⁻⁵	3.0 x10 ⁻³
	MF	0005515	protein binding	3.0 x10 ⁻⁴	8.3 x10 ⁻³
	MF	0016787	hydrolase activity	3.1 x10 ⁻³	4.1 x10 ⁻²
	BP	0007160	cell-matrix adhesion	1.9 x10 ⁻⁴	4.0 x10 ⁻²
	CC	0044459	plasma membrane part	2.2 x10 ⁻⁴	1.5 x10 ⁻²
	BP	0009966	regulation of signal transduction	1.1 x10 ⁻⁴	3.2 x10 ⁻²
	MF	0004713	protein tyrosine kinase activity	1.9 x10 ⁻³	3.5 x10 ⁻²

3.4 Discussion

3.4.1 Scope of the study

By looking at the evolution of multi-domain proteins, I address here the question of mechanisms of creation of novel animal genes. The current state in the field is that the approach to this problem is more theoretical and centers around the rare clear examples of novel gene creation (Long, 2001). This is the first study that systematically looked at the mechanisms that created novel, more complex, animal genes. My approach to this was to present proteins as strings of functional domains and look at the domain rearrangements. Earlier studies that examined characteristics of gained or lost protein domains were comparing proteins with similar domain architectures, which alone did not allow distinction between gain and loss events (Bjorklund et al., 2005; Weiner et al., 2006). Here, I use direct phylogenetic relations among animal genes to identify a high-confidence set of protein domain gain events, which enabled me to study general trends in evolution of more complex domain architectures in the animal kingdom. Secondly, I relate information from the proteins to the underlying exon structures to help elucidate the causative mechanisms. To assign domains to proteins, I used Pfam-A domain annotations. However, Pfam-A is not comprehensive, and inclusion of unassigned regions could have increased the number of inferred domain gains in the study. Additionally, profile HMMs for individual Pfam domains do not necessarily cover all related sequences. I have tried to overcome this by grouping domains into clans, which include more distantly evolutionarily related domain profiles. However, even after domain refinements, it is possible that domain assignments are sometimes falsely omitted from the sequences. To avoid false domain gain calls, I excluded all similar sequences that differed in domain assignments from the analysis (Section 3.2.2). This again lowered the number of inferred domain gain events. The main aim of this study was to obtain a set of high confidence domain gain events. However, by excluding possible false cases of domain gain events, real cases might have been missed too.

To find a set of high confidence domain gain events, I used gene phylogenies of completely sequenced animal genomes from the TreeFam database (Ruan et al., 2008). TreeFam contains phylogenetic trees of animal gene families, and is able to assign ortholog and paralog relationships because it records the positions of speciation and duplication events in the phylogenies. I assigned domains to the protein sequences in these families according to Pfam annotation (Finn et al., 2008). The Pfam database provides the most comprehensive collection of manually curated protein domain signatures. Its family assignments are based on evolutionarily conserved motifs in the protein sequences.

3.4.2 Approach for obtaining the set of confident domain gain events

The relative frequencies of domain gain and loss events are not known and most probably not universal for different domains and organisms. Hence, different approaches have been undertaken to address this issue. Several previous studies have assumed that the frequency of gain and loss events are equal and have identified domain gains and losses by applying maximum parsimony (Kummerfeld and Teichmann, 2005); (Buljan and Bateman, 2009; Fong et al., 2007; Forslund et al., 2008). Other studies have assumed that domain loss is slightly more likely than domain gain (Itoh et al., 2007) or that the difference in the frequency of gains and losses is very significant and hence have suggested Dollo parsimony (which allows a maximum of one gain per tree) for identifying domain gains (Basu et al., 2008; Przytycka et al., 2006). I found that the set of domain gains obtained by applying maximum parsimony was heavily enriched in cases that were misidentified multiple domain losses in the tree. Therefore, it is also possible that the frequency of gene fusions and reinvention of domain architectures is smaller than previously proposed (Kummerfeld and Teichmann, 2005; Fong et al., 2007; Forslund et al., 2008). On the other hand, if there were situations where the same domain was gained more than once in the same gene family, Dollo parsimony would still predict only one domain gain and would not distinguish different gain events. Therefore, my approach was to identify domain

gains by assuming that the losses were slightly more likely than gains (by applying Weighted parsimony) and then filter these to only include trees with a single gain (using the rationale of Dollo parsimony). This strategy appeared to reduce the number of likely false domain gains as judged by inspection of the results.

3.4.3 Mechanisms of domain gain

Present domain combinations are shaped by the causative molecular mutation mechanisms followed by natural selection. In this chapter, I addressed the question of what mechanisms have been and possibly still are creating novel, more complex, animal domain architectures and hence new functional arrangements. I investigated the supporting evidence for the mechanisms that are believed to be candidates for the observed domain gains and found several examples of domain gain that can be clearly connected with their causal mechanisms. These examples illustrate domain gain through retroposition and through joining of exons from adjacent genes.

The SETMAR gene, an example for the role of retroposition, is of particular interest because it adds to the list of only a few known examples of novel gene creation in the human lineage assisted by this mechanism. It was discussed before that retroposed domains are most likely to be found at the C-termini of genes (Babushok et al., 2007b). By this means, the issue of transcription regulation would be avoided. In the case of the SETMAR gene, the retroposed domains are at the N-terminus. However, this gene lies in the intron of another gene on the opposite strand. This suggests that transcription of the SETMAR gene could be facilitated by open chromatin structure and transcription of the gene that it overlaps with. Interestingly, a similar phenomenon was reported for the novel human genes that evolved from noncoding DNA (Knowles and McLysaght, 2009). A lack of evidence for other candidate cases is not a definite proof that retroposition was not the active mechanism. Frequency of multi-exon domains is higher among the ‘ancestral’ domains in the representative proteins, i.e. among those domains that were not categorized as

gained domains in this study (86% of the 'ancestral' domains in the representative proteins are encoded by two or more exons, Section 3.3.2). This could imply that domains encoded by a single exon were more easily inserted into proteins during evolution, or even that among the gained domains are other cases of domain retroposition. In addition, intron insertions during evolution of animal genes could have camouflaged the cases of domain gains through retroposition. However, more than 70% of the gained domains in the whole set are encoded by more than one exon, and extra exons have also frequently been gained together with the gained domains which are encoded by a single exon (Section 3.3.6.2). Intron presence in the majority of the gained domains would therefore suggest that retroposition did not have a major role in the evolution of animal domain architectures.

With regard to other lineages, only the gains in insects, with representative proteins from *Drosophila melanogaster*, have numerous examples (22 cases) of a gain of domain coded by one exon, leaving open the possibility that retroposition might be a more important mechanism for domain gain in insects than it is in other lineages. However, overall this seems to be a rare mechanism for domain gain in animals. Additionally, it is important to note that previous work also underlined the role of adjacent gene joining (Zhou et al., 2008) and NAHR (Yang et al., 2008) in the formation of chimeric genes in the *Drosophila* lineage.

The dominant mechanism for domain gains in the animal genomes appears to be joining of exons from adjacent genes. Additionally, this mechanism seems to be in a strong connection with gene duplication. Apart from showing here the evidence for the dominant role of adjacent genes' exons joining, I also find the examples that directly illustrate how this mechanism operates. These examples are shown in Figure 3.8. After duplication, exons that encode one or more domains are joined with exons from an adjacent gene. The examples are interesting from the point of view of evolution of protein diversity, but also as additional examples for novel gene creation during animal evolution. In addition, I addressed here the possible role of recent segmental duplications in gene evolution. As a result, I found two genes that were created after a segmental duplication event. The possible mechanism for creation of these genes is

illustrated in Figure 3.10. However, it is necessary to be cautious when assessing the possible roles of these proteins. For both examples, there is only transcript evidence and some of the transcript products of these genes appear to have a structure that would lead to them being targeted by NMD (Wilming et al., 2008). Sometimes it is possible for a transcript to avoid the NMD signal and in this case these examples would be of high interest as possible sources of novel function. In the case that these transcripts are silenced by NMD, these genes are still interesting examples from the theoretical point of view; they directly illustrate the mechanism of how gene evolution can work. Initially, part of a gene sequence gets duplicated and recombined with another gene; if juxtaposed exons are in frame, a joint transcript can be created and through NMD deleterious protein variants can be silenced at the transcript level while allowing at the same time introduction of novel mutations that can be tested later on.

Another mechanism that can cause gain of a novel protein domain is exonisation of a previously non-coding sequence. Here, I observe that domains which are gained as exon extensions are preferentially disordered (Figure 3.11). If a new protein domain is gained from a previously non-coding sequence it is more likely that the encoded protein region will not be structured and that the sequence will be inserted through exon extension rather than as a completely new exon. Hence, disordered protein regions, which are gained as exon extensions are likely candidates for a domain gain through exonisation of non-coding sequence. Conversely, this also suggests a possible mechanism for evolution of disordered protein regions. An illustration from the literature for the significance of inclusion of novel disordered segments into proteins is the evolution of NMDA receptors. These receptors display a vertebrate specific elongation at the C-terminus. Gained protein regions are disordered and govern novel protein interactions, and it is believed that this might have contributed to evolution and organization of postsynaptic signalling complexes in vertebrates (Ryan et al., 2008).

Further support for the assumption that domain gains through exon extensions are enriched in gains caused by exonisation of previously non-coding sequences comes from the observed bias for these gains to occur at the C-terminus (Figure 3.5). Namely, it is expected that gains by exonisation are most

likely to be observed at C-terminus since extension of exons at N-terminus or in the middle of proteins can introduce frame shifts and hence can be selected against. However, Pfam families that are classified as exon extensions are also likely to be shorter so it is possible that this introduces some bias, since shorter families are less likely to be domains with defined structures. Moreover, an important caveat is that only a systematic study can confirm domain gain by this mechanism; apparently non-coding sequences, which are homologous to gained domains, might only lack transcript and protein evidence in the less studied species and thus miss domain assignment. In addition, it is important to note that exonisation of previously noncoding sequences is not the only mechanism that can explain exon extensions. Other possible mechanisms are gene recombination inside exon regions and deletion of sequences between exons of two adjacent genes.

Analysis of the high confidence set of domain gain events suggests that retroposition and recombination-assisted intronic insertions, in contrast to previous expectations (Kaessmann et al., 2002; Liu and Grigoriev, 2004), are minor contributors to domain gains. Therefore, it is possible that the role of intronic insertions had been overestimated previously. It will be interesting to see if the observed excess of symmetrical intron phases around exons coding for domains (Kaessmann et al., 2002) is due to exon shuffling or to some other mechanism such as selective pressure from alternative splicing (Lynch, 2002).

3.4.4 Domain gains were assisted by recombination events

Gained domains can have an origin in the neighboring genes or non-coding sequences, or they can be inserted into another gene by the transposon machinery. Results presented in this chapter suggest that exonisation of non-coding sequence and retroposition were not the mechanisms that caused the majority of the high confidence gain events. Additionally, the analysis showed that in animals without the reported gain, genes homologous to those whose exons were joined together were not adjacent to each other on the genome.

Hence, the most probable explanation is that the majority of these events were preceded by recombination, which juxtaposed novel gene combinations.

In 80% of the gain events, a domain gain has occurred after duplication of either a 'donor' or 'acceptor' gene. Retroposition does not seem to be a valid explanation for the majority of these duplications and it is possible that they were created by a recombination mechanism. Additionally, I observed a bias in the chromosomal positions of the plausible 'donor genes' in the way that they were preferentially found on the same chromosomes as genes with the gained domains. The bias was more prominent for the younger gain events (Figure 3.12), possibly due to continuous chromosomal rearrangements. NAHR creates duplicates more frequently than IR does (Freeman et al., 2006; Roth et al., 1985), creates them preferentially on the same chromosome (Freeman et al., 2006) and provides ground for gene rearrangements. Therefore, it is possible that NAHR assisted domain gains, and in particular preceded joining of exons from adjacent genes. I do not exclude IR as a possible causative mechanism but NAHR seems more likely given the bias in chromosome locations of domain duplicates and reliance of the gain mechanism on gene duplication. Moreover, recent work by Kim and colleagues (Kim et al., 2008) has suggested that even though IR might be important for the formation of new copy number variants in the human genome, NAHR - mediated by Alu elements and existing segmental duplications themselves - had a dominant role in the formation of fixed segmental duplicates.

If recombination acted to juxtapose novel domain combinations, it is possible that it directly created novel introns and joined exons from the two adjacent genes. However, it is more likely that recombination only brought novel exons from two different genes into proximity, allowing alternative splicing to create novel splice variants. As discussed above, there are indications that NAHR could have caused the initial duplications and rearrangements. The implications for the role of NAHR in animal evolution in general are particularly interesting since this mechanism is still primarily associated with more recent mutations in the human genome (and primate genomes in general), structural variations in human population and disease development (Bailey and Eichler, 2006; Conrad and Hurles, 2007; Stankiewicz and Lupski, 2002). It has, however, recently been proposed that other mechanisms, such as Fork Stalling and Template Switching

(FoSTeS) mechanisms could have also had a role in genome and single-gene evolution. FoSTeS (Zhang et al., 2009), a replicative mechanism that relies on microhomology regions, seems to provide a better explanation for complex germline rearrangements, but also for some tandem duplications in the genome, than NAHR and IR (Gu et al., 2008). Hence, the exact relative contributions of these different mechanisms are still to be determined. However, this might be hampered by sequence divergence after domain gain events, which have occurred millions years ago.

In conclusion, work presented in this chapter gives evidence for the importance of gene duplication followed by adjacent gene joining in creating genes with novel domain-combinations. The role of duplicated genes in donating domains to adjacent proteins is a potentially important, and powerful, mechanism for neofunctionalisation of genes.

3.4.5 Different trends in domain gains in different lineages and at different time points during evolution

It is important to note that even though I have attempted here to draw general conclusions about dominant mechanisms for evolution of animal genes, it is possible that contributions of different mechanisms differ between different species and at different time points during evolution. The percentage of active retrotransposons, rates of chromosomal rearrangements and intergenic splicing can be different in different genomes. Similarly, selection force, which decides on toleration of intermediate stages in gene evolution, depends on the population size and will differ between different species. Therefore, it is possible that we will find evidence that some mechanisms are more relevant in some species than they are in others. This is illustrated with differences in characteristics of the gained domains in vertebrates and *Drosophila*. The dominant mechanism in *Drosophila* seems to be the extension of exons at the C-terminus. Additionally, even though the majority of gain events are represented by human proteins, different mechanisms could have dominated at different evolutionary time points in the human lineage. For example, LINE-1 retrotransposons are abundant in mammals but not in other animals (Han and Boeke, 2005), and

whole genome duplication that occurred after divergence of vertebrates (Dehal and Boore, 2005) could have preferred recombination between gene duplicates at that point in time.

3.4.6 Functional implications of domain gain events

Creation of novel genes is assumed to play a crucial role in the evolution of complexity. Previous studies have put a considerable effort into identifying gene gain and loss events during animal evolution, as well as into analyzing functional and expression characteristics of these genes (Blomme et al., 2006; Hahn et al., 2007; Milinkovitch et al., 2010; Tzika et al., 2008). In this study, my aim was to investigate functionally relevant changes of individual proteins. Implications of observed domain gains on the evolution of more complex animal traits are highlighted by the frequent regulatory function of the gained domains in the human lineage. Shuffling of regulatory domains has already been proposed as an important driving force in the evolution of animal complexity (Peisajovich et al., 2010; Pawson and Nash, 2003), and an increase in the number of regulatory domains in the proteome has been directly related to the increase of organism complexity (Vogel and Chothia, 2006).

3.5 Bibliography

- Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A., and Sorek, R. (2006). Transcription-mediated gene fusion in the human genome. *Genome research* 16, 30-36.
- Arguello, J.R., Fan, C., Wang, W., and Long, M. (2007). Origination of Chimeric Genes through DNA-Level Recombination. *Genome Dyn* 3, 131-146.
- Babushok, D.V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Mandal, P.K., Okada, N., Abrams, C.S., and Kazazian, H.H., Jr. (2007a). A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome research* 17, 1129-1138.
- Babushok, D.V., Ostertag, E.M., and Kazazian, H.H., Jr. (2007b). Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci* 64, 542-554.
- Bailey, J.A., and Eichler, E.E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature reviews* 7, 552-564.
- Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. *Science (New York, NY)* 297, 1003-1007.
- Basu, M.K., Carmel, L., Rogozin, I.B., and Koonin, E.V. (2008). Evolution of protein domain promiscuity in eukaryotes. *Genome research* 18, 449-461.
- Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J., and Elofsson, A. (2005). Domain rearrangements in protein evolution. *Journal of molecular biology* 353, 911-923.
- Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome biology* 7, R43.
- Buljan, M., and Bateman, A. (2009). The evolution of protein domain families. *Biochemical Society transactions* 37, 751-755.
- Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science (New York, NY)* 300, 1701-1703.

- Conrad, D.F., and Hurles, M.E. (2007). The population genetics of structural variation. *Nature genetics* 39, S30-36.
- Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proceedings of the National Academy of Sciences of the United States of America* 103, 8101-8106.
- Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology* 3, e314.
- Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology* 4, P3.
- Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology* 347, 827-839.
- Ekman, D., Bjorklund, A.K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of molecular biology* 372, 1337-1348.
- Farris, J.S. (1977). Phylogenetic analysis under Dollo's Law. *Systematic Zoology* 26, 77-88.
- Finn, R.D., Tate, J., Mistry, J., Coghill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., et al. (2008). The Pfam protein families database. *Nucleic Acids Res* 36, D281-288.
- Fong, J.H., Geer, L.Y., Panchenko, A.R., and Bryant, S.H. (2007). Modeling the evolution of protein domain architectures using maximum parsimony. *Journal of molecular biology* 366, 307-315.
- Forslund, K., Henricson, A., Hollich, V., and Sonnhammer, E.L. (2008). Domain tree-based analysis of protein architecture evolution. *Molecular biology and evolution* 25, 254-264.
- Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. *Genome research* 16, 949-961.

- Gilbert, W. (1978). Why genes in pieces? *Nature* 271, 501.
- Gsponer, J., and Babu, M.M. (2009). The rules of disorder or why disorder rules. *Progress in biophysics and molecular biology* 99, 94-103.
- Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science (New York, NY)* 322, 1365-1368.
- Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. *Pathogenetics* 1, 4.
- Hahn, M.W., Demuth, J.P., and Han, S.G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics* 177, 1941-1949.
- Han, J.S., and Boeke, J.D. (2005). LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression? *Bioessays* 27, 775-784.
- Itoh, M., Nacher, J.C., Kuma, K., Goto, S., and Kanehisa, M. (2007). Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome biology* 8, R121.
- Kaessmann, H., Zollner, S., Nekrutenko, A., and Li, W.H. (2002). Signatures of domain shuffling in the human genome. *Genome research* 12, 1642-1650.
- Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnam, N.H., Rokhsar, D.S., Kanehisa, M., Satoh, N., and Wada, H. (2009). Domain shuffling and the evolution of vertebrates. *Genome research* 19, 1393-1403.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research* 12, 996-1006.
- Kim, P.M., Lam, H.Y., Urban, A.E., Korb, J.O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., and Gerstein, M.B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome research* 18, 1865-1874.
- Knowles, D.G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome research* 19, 1752-1759.
- Kummerfeld, S.K., and Teichmann, S.A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet* 21, 25-30.

- Liu, M., and Grigoriev, A. (2004). Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet* 20, 399-403.
- Liu, M., Walch, H., Wu, S., and Grigoriev, A. (2005). Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic acids research* 33, 95-105.
- Long, M. (2001). Evolution of novel genes. *Curr Opin Genet Dev* 11, 673-680.
- Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature reviews* 4, 865-875.
- Long, M., Rosenberg, C., and Gilbert, W. (1995). Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Sciences of the United States of America* 92, 12495-12499.
- Lynch, M. (2002). Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences of the United States of America* 99, 6118-6123.
- Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics* 24, 2630-2631.
- Magrangeas, F., Pitiot, G., Dubois, S., Bragado-Nilsson, E., Cherel, M., Jobert, S., Lebeau, B., Boisteau, O., Lethe, B., Mallet, J., et al. (1998). Cotranscription and intergenic splicing of human galactose-1-phosphate uridylyltransferase and interleukin-11 receptor alpha-chain genes generate a fusion mRNA in normal cells. Implication for the production of multidomain proteins during evolution. *The Journal of biological chemistry* 273, 16005-16010.
- Milinkovitch, M.C., Helaers, R., and Tzika, A.C. (2010). Historical constraints on vertebrate genome evolution. *Genome Biol Evol* 2010, 13-18.
- Nurminsky, D.I., Nurminskaya, M.V., De Aguiar, D., and Hartl, D.L. (1998). Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396, 572-575.
- Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E., and Guigo, R. (2006). Tandem chimerism as a means to increase protein complexity in the human genome. *Genome research* 16, 37-44.

- Patthy, L. (1996). Exon shuffling and other ways of module exchange. *Matrix Biol* 15, 301-310; discussion 311-302.
- Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling--a review. *Gene* 238, 103-114.
- Patthy, L. (2008). Exons and Protein Modules. In *Encyclopedia of life sciences* (John Wiley & Sons, Ltd.).
- Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science (New York, NY)* 300, 445-452.
- Peisajovich, S.G., Garbarino, J.E., Wei, P., and Lim, W.A. (2010). Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science (New York, NY)* 328, 368-372.
- Ponting, C.P. (2008). The functional repertoires of metazoan genomes. *Nature reviews* 9, 689-698.
- Przytycka, T., Davis, G., Song, N., and Durand, D. (2006). Graph theoretical insights into evolution of multidomain proteins. *J Comput Biol* 13, 351-363.
- Roth, D.B., Porter, T.N., and Wilson, J.H. (1985). Mechanisms of nonhomologous recombination in mammalian cells. *Molecular and cellular biology* 5, 2599-2607.
- Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L.J., Guo, Y., Heriche, J.K., Hu, Y., Kristiansen, K., Li, R., et al. (2008). TreeFam: 2008 Update. *Nucleic Acids Res* 36, D735-740.
- Ryan, T.J., Emes, R.D., Grant, S.G., and Komiyama, N.H. (2008). Evolution of NMDA receptor cytoplasmic interaction domains: implications for organisation of synaptic signalling complexes. *BMC neuroscience* 9, 6.
- Sankoff, D., Cedergren, R.J., and McKay, W. (1982). A strategy for sequence phylogeny research. *Nucleic Acids Res* 10, 421-431.
- Stankiewicz, P., and Lupski, J.R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet* 18, 74-82.
- Thomson, T.M., Lozano, J.J., Loukili, N., Carrio, R., Serras, F., Cormand, B., Valeri, M., Diaz, V.M., Abril, J., Burset, M., et al. (2000). Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Res* 10, 1743-1756.

- Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S., and Hurles, M.E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* 40, 90-95.
- Tzika, A.C., Helaers, R., Van de Peer, Y., and Milinkovitch, M.C. (2008). MANTIS: a phylogenetic framework for multi-species genome comparisons. *Bioinformatics (Oxford, England)* 24, 151-157.
- van Rijk, A., and Bloemendal, H. (2003). Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica* 118, 245-249.
- Vogel, C., and Chothia, C. (2006). Protein family expansions and biological complexity. *PLoS computational biology* 2, e48.
- Weiner, J., 3rd, Beaussart, F., and Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *FEBS J* 273, 2037-2047.
- Wilming, L.G., Gilbert, J.G., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J.L. (2008). The vertebrate genome annotation (Vega) database. *Nucleic acids research* 36, D753-760.
- Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology* 293, 321-331.
- Yang, S., Arguello, J.R., Li, X., Ding, Y., Zhou, Q., Chen, Y., Zhang, Y., Zhao, R., Brunet, F., Peng, L., et al. (2008). Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS genetics* 4, e3.
- Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature genetics* 41, 849-853.
- Zhang, X.H., and Chasin, L.A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proceedings of the National Academy of Sciences of the United States of America* 103, 13427-13432.
- Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in *Drosophila*. *Genome research* 18, 1446-1455.