Chapter 2

Evolution of multidomain proteins

2.1 Introduction

In this chapter, I investigate the general trends of protein domain architecture evolution. To decrease the number of falsely reported domain gain and loss events, I first develop a method for the refinement of initial domain assignments. Next, I analyse the positions in proteins where the changes in domain architectures are reported. Positions of changes are defined by the mechanism that caused domain gains or losses and by subsequent natural selection. Here, I analyse the differences in trends between the changes that occurred after gene duplication or organism speciation and a possible role of natural selection in this.

Protein domains, as defined here, are conserved regions of a protein's sequence that often convey distinct function. The domain architecture, or order of domains in a protein, is considered as a fundamental level of protein functional complexity (Holm and Sander, 1994) and assignment of domains to a protein is an important step in elucidation of a protein's function (Bateman et al., 2002). The majority of the protein repertoire is composed of multidomain proteins; two-thirds of the proteins in prokaryotes and about four-fifths eukaryotic proteins have two or more domains (Chothia et al., 2003). Moreover, an organism's complexity relates much better to the number of distinct domain architectures (Babushok et al., 2007) and expansion in particular domain

families (Vogel and Chothia, 2006) than to the number of genes in the organism. The prevalence of proteins with more than two domains and the recurrent appearance of the same domain in otherwise non-homologous proteins show that functional domains are reused when creating new proteins. Because of this, domains have been likened to Lego bricks that can be recombined in various ways to build proteins with completely new functions (Das and Smith, 2000). Hence, one way to study the evolution of protein function and structure is by looking at the evolution of protein domain architecture. The average length of a protein domain is around 120 amino acids, so changes in domain architecture are in general underlined by large alterations at the gene level.

Good quality domain annotations of proteins are important for better understanding of protein evolution and function. However, they are also a necessary pre-requirement for studies that aim to address the evolution of protein domain architecture. Domain prediction methods have successfully applied profile hidden Markov models (HMMs) for identifying protein domains within amino acid sequences (Bateman et al., 2000). Nonetheless, these methods are still not able to successfully predict all domains in proteins and the missing domain assignments could assist in explaining protein function. There have been several attempts to improve domain annotation of proteins. For example, the speech recognition techniques that rely on the usage of language modelling have been adapted to find domains in protein sequences (Coin et al., 2003). The reasoning behind this approach is that certain word, or domain, combinations are more likely than others and hence domain detection relies on context, i.e. the presence of other domains in a protein (Coin et al., 2003). Similarly, information about the taxonomic distribution of domains has been incorporated into domain recognition algorithm, which also resulted in the enhanced domain recognition (Coin et al., 2004). The two latter approaches have been applied to increase the coverage of proteins with Pfam assignments. Context analysis has also been used to add missing domains to proteins that had a highly similar domain architecture and sequence similarity in the region that had an extra domain assigned to one of the compared proteins only (Beaussart et al., 2007). However, the latter method, named AIDAN, has so far been done only for proteins with more than six domains and domain assignments from the ProDom database (Beaussart et al.,

2007). The ProDom database (Bru et al., 2005) uses recursive PSI-Blast search for domain annotation and has a lower coverage than the Pfam database.

Previous studies have been addressing the evolution of novel domain architectures by comparing homologues with similar domains and investigating positions in proteins where the changes occurred. By doing this, the authors were able to give predictions about the mechanisms that caused the observed rearrangements. Among the molecular mechanisms that can direct protein rearrangements are gene fusion and fission (Moore et al., 2008), exon shuffling through intronic recombination (Patthy, 1999), alternative gene splicing, introduction of novel stop codons and retroposition (Babushok et al., 2007). In prokaryotes, gene fusion and fission are reported to be the major drivers of changes in protein domain composition (Enright et al., 1999; Pasek et al., 2006). However, little is still known about exact mechanisms that underlie these changes in eukaryotes (Babushok et al., 2007; Moore et al., 2008). A study by Weiner *et al.* reported that changes in domain architecture preferentially occur at the protein termini, which was in agreement with previous reports (Bjorklund et al., 2005). In their study, Weiner *et al.* assumed that the frequency of domain deletions is much higher than the frequency of domain insertions and proposed that introductions of novel start and stop codons are the major causative mechanisms for changes in domain architectures (Weiner et al., 2006).

A special aspect of the evolution of protein domain architectures is the evolution of protein domain repeats; the difference between a gain and loss of a single copy domain and a tandemly repeated domain in a repeat is illustrated in Figure 2.1. Many proteins, especially in eukaryotes, contain tandem copies of the same domain (Bjorklund et al., 2006). Mechanisms that have governed changes in the number of domain repeats are not well understood, and they are not necessarily the same as the ones that have directed gains and losses of single copy protein domains. In fact, Bjorklund *et al.* found that many of the repeats have been duplicated in the middle of the repeat region (Bjorklund et al., 2006). Expansion of domain repeats is important for the evolution of protein function; domain repeats have a variety of binding functions and proteins with them tend to have more interaction partners in protein-protein interaction networks than those without (Ekman et al., 2005). An interesting illustration for the important

functional role played by domain repeats is in the gene *Prdm9*. Mouse *Prdm9* encodes a protein with a KRAB motif, a histone methyltransferase domain and several zinc fingers. A difference in the number of zinc finger repeats is a trait that distinguishes alleles which cause hybrid sterility from those that do not (Oliver et al., 2009).

Apart from being reliant on the mechanisms that create them, existing domain combinations are also a result of selective forces that enabled them to remain in a population. Selective forces, which act on proteins, depend, among other factors, on the evolutionary pressure to preserve the original protein function as it was. This could be relieved when the changes in domain architecture follow gene duplication and one copy can freely evolve while the other stays intact. Furthermore, a pressure to remove a protein from a population also depends on how the overall protein function is affected by domain gain or loss. For example, whether domain loss leads to protein subfunctionalization or completely abolishes the original function, and similarly, when a domain is gained - whether the function of the gained domain is compatible with the function, or localization, of other domains in the ancestral protein. Finally, structural stability of a novel protein is also a crucial factor which determines whether the new domain architecture will be preserved or not. Interestingly, some domains are observed in a number of different domain combinations, and are considered to be 'promiscuous', whereas others occur in only one or a few combinations (Marcotte et al., 1999). These promiscuous domains are, typically, involved in protein–protein interactions, and some of them play important roles in signalling pathways (Basu et al., 2008). This, together with the fact that they show evidence of strong purifying selection acting on them (Basu et al., 2008), implies that these domains were able to become promiscuous in the first place because they had a potential to be useful in various contexts.

Evolution of protein domain architectures has so far been addressed in a number of studies. However, there is no agreement in the field on what is the relative frequency of domain gain and loss events. In particular, there were different reports on the rate of convergent evolution of domain architectures (Forslund et al., 2008; Gough, 2005). Furthermore, depending on the study,

changes in domain architectures were interpreted predominately as a result of domain gains (Bjorklund et al., 2005) or of domain losses (Weiner et al., 2006). Similarly, different algorithms were applied to find domain gains and losses. Some of these approaches assumed domain gain and loss to be equally likely (Fong et al., 2007; Forslund et al., 2008; Kummerfeld and Teichmann, 2005), while other considered domain loss to be a more likely event than domain gain (Basu et al., 2008; Itoh et al., 2007).

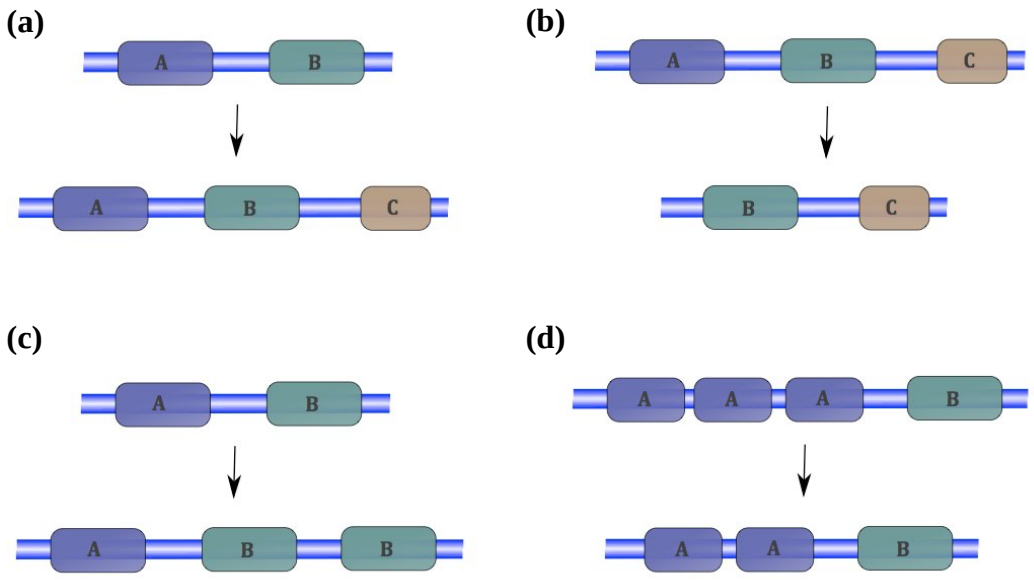

Figure 2.1 Illustration of domain gains and losses. Figure (a) illustrates gain of a novel domain and figure (b) loss of a domain, which was present in one copy in the ancestral protein. Figure (c) illustrates a domain gain, which leads to a domain repeat and figure (d) loss of a domain from a repeat.

2.2 Methods

2.2.1 Analysis of TreeFam families

The TreeFam database provides information about phylogenetic trees of animal gene families. TreeFam infers orthology by fitting a gene tree into a universal species tree and finds historical duplication, speciation and gene loss events (Li et al., 2006). The database has a very good coverage of fully sequenced animal genomes, including for example 84.5% of known human protein-coding genes. It consists of two parts; gene families whose trees have been manually curated, termed TreeFam-A, and those that have only automatically created trees, termed TreeFam-B. Genes in the TreeFam-A families are of better quality but are, for example, biased to those involved in mitotic processes. Therefore, to have a comprehensive view of trends in domain architecture evolution I included both TreeFam-A (1,305) and TreeFam-B (14,345) gene families in the analysis (TreeFam release 4.0). To infer relations among genes in a family, I used each family's clean tree. Clean trees contain genes from 25 fully sequenced animal genomes, together with yeast and plant outgroups. For parsing trees, I used the TreeFam API (http://treesoft.sourceforge.net/). Genes in TreeFam trees are represented with transcripts that are most similar to other transcripts in the tree.

2.2.2 Assignment of domains to proteins with refinement

I assigned Pfam-A domains (release 22.0) to all protein products of TreeFam transcripts using the Pfam_scan.pl software. Since domains in the same Pfam clan are evolutionary related, I replaced domain identifiers with clan identifiers where applicable. Domain prediction methods can both fail to predict bona fide domains as well as make false predictions, which look like domain losses and gains respectively. To address this issue, I applied a refinement process; I firstly removed the likely false positive fragmentary domain assignments, i.e. domains that were called on only a single sequence in a TreeFam family with an E-value

larger than 10^{-6} and only 30% or less of the domain's Pfam model covered. Next, when some sequences lacked a domain, which was annotated to other family members, I used Wu-blastp to search the domain sequence against the protein sequences not annotated with the domain. When a significant match was found (E-value less than 10^{-4} and at least 60% of a domain sequence present, or alternatively an E-value less than 10^{-7} and 40% or more of a domain sequence present, or only E-value less than 10^{-10} and any length of the matched sequences) I added domain assignments to the sequences. I iterated the procedure for all newly assigned domains until no new domain assignments were found.

2.2.3 Domain gains and losses

To identify domain gain and loss events, I applied the maximum parsimony algorithm. The rationale behind the algorithm is that the evolutionary scenario explained with as few events as possible is the most probable one. The algorithm firstly infers domain composition of ancestral sequences in the trees and then compares the ancestral with their daughter sequences. To record the position of changes in proteins - i.e. N-, C-terminal or middle - I implemented the Needleman-Wunsch algorithm, which aligned proteins as strings of domains. When changes in the domain architectures could have been explained with gains or losses of domains at different positions, I reported the inferred gain or loss for each of these positions, but multiplied it with the likelihood of the scenario. For example, when a domain repeat at the termini expanded, I assigned the change as both - possible domain insertion at the termini and possible insertion in the middle of a protein, with the probability for each scenario depending on the number of domains in the ancestral repeat.

To calculate the expected number of domain gains and losses at each position, I took into account the domain composition of ancestral proteins that experienced changes in domain architecture. I assumed that domain gain or loss is equally likely to occur at the N-termini, C-termini or in the middle of a protein. Hence, an ancestral protein with three domains is assumed to have equal probability of losing a domain at any position, but for an ancestral protein with four domains, which then has two middle domains, there is 50% probability that

a lost domain will be from the middle of a protein. Similarly, an ancestral protein with two domains is assumed to be equally likely to gain a domain at any position, but the ancestral protein with three domains has two positions where a new domain could be inserted as a middle domain and hence 50% probability that a domain gain will occur in the middle of a protein. The total number of expected changes at each position is calculated by adding the expected number of changes for the ancestral proteins of each length. This is obtained by multiplying the probability of the change at each position with a total number of gains or losses observed for ancestral proteins with a given number of domains. Positions of changes were not defined for ambiguous events where domains were added to ancestral sequences with no domains and where all domains from ancestral sequences were lost. Statistical significance of the observed trends was assessed with the R software.

The costs for domain gain and loss in the maximum parsimony algorithm are equal. However, to investigate how a starting assumption about the frequency of one event over another influences the ratio of reported domain gain and loss events, I implemented a weighted parsimony algorithm. By changing the relative costs of domain gain and loss events in the algorithm, one changes the assumption about the relative frequency of these events. I studied how the ratio of reported events depends on the input parameters of the algorithm.

The approach in this study was to infer domain architectures of the ancestral proteins by looking at the domain composition of present day proteins. However, after species divergence or gene duplication, homologous proteins evolve at different rates and neither of them necessarily maintains the ancestral domain composition. Therefore, the inferred domain gain and loss events do not include all possible scenarios. Also, in the cases where neither of the descendants has a domain that was present in the ancestral protein, its domain composition cannot be correctly reconstructed by this approach.

2.3 Results

2.3.1 Phlyogenetic trees can guide refinement of domain assignments

In order to improve the quality of domain annotations for the proteins in the TreeFam database, I made use of their inferred phylogenetic relations. When there were inconsistencies in domain assignments between the members of the same TreeFam family, I analysed their protein alignments and refined the initial domain assignments when this was justifiable. If only one member of a gene family had a domain annotated to it; I noted the probability with which this domain was assigned, and the fraction of an HMM model for the domain that was mapped to a motif in the sequence. If these were not significant (see Methods section 2.2.2), the annotation was considered as a false positive. This procedure detected 115 false positive domain assignments in all TreeFam proteins (listed in Appendix A.1). These matches were reported to the Pfam database so that their family thresholds could be redefined and the false positive hits removed. For all other inconsistencies in domain annotation, I analysed whether a domain assignment was falsely missing from the proteins that lacked the annotation present in their homologues. When sequence similarity between the aligned protein regions which differed in domain annotation was significant, domain annotations were added to the sequences missing them. To look for similarity, I used Wu-blastp, which is a faster procedure than using a profile-HMM. However, Wu-blastp does not take into account conservation of different amino acids in a motif and is not as sensitive as a profile-HMM. To assess its suitability for refinement of domain assignments I performed a test where in each TreeFam family I deleted Pfam domain assignments in all but one protein and then investigated how well these could be recovered with the refinement algorithm. For this, I randomly selected 100 TreeFam families and repeated the analysis 10 times on different sets of families. I found that on average this procedure recovered 95% of the initial domain assignments. This is likely an overestimate since domains that were recovered were initially predicted and because of that, are potentially more significantly similar to the model and hence to each other.

Nevertheless, this showed that Wublastp with the criteria described in Methods could be used for adding erroneously missing domain assignments. At least one missing domain was added to 15% of all TreeFam proteins. This increased both sequence coverage - i.e. percentage of proteins with at least one domain assigned to them - by 5%, and residue coverage - i.e. percentage of all residues covered with Pfam domains - by 10% of the proteins. Residue and sequence coverage of the TreeFam proteins before and after domain refinements is shown in Table 2.1. Finally, TreeFam families that lacked any domain assignment are interesting from the point of view of identification of novel protein domains. There were 4,445 gene families, out of total 15,656 TreeFam-A and -B families, that lacked any domain assignment. I reported these families to the Pfam database so that the shared homologues sequences in them could be used for building of new Pfam families. All these gene families belonged to TreeFam-B and many of them contained only a few protein sequences. Hence, the most interesting here are those families with many homologous sequences but no known domain assignment; 1,181 TreeFam families had ten or more genes and no domain annotation for any of them.

Success in annotating domains to proteins depends on how well a model for each domain represents the domain and how specific it is for a particular domain. This is likely to be strongly influenced by the sequence content and length of each domain. I have looked at how the quality of domain predictions in TreeFam proteins depends on the length of domain models. Quality of domain predictions is represented with the consistency of domain assignments between proteins that belong to a same TreeFam-A family, i.e. between proteins that are with high confidence grouped together in a gene family. I have found that with shorter domains, there is more inconsistency in assignments of domains (Figure 2.2). In particular, domains for which models are shorter than 50 amino acids are on average predicted in only half of the proteins in a phylogenetic tree. Inconsistency of annotations is partly due to real domain gains and losses. However, a strong bias for the quality of annotations to be correlated with the length of domain models confirms an expectation that the shorter the domain model is, the more difficult it is to get a significant score for the presence of the motif in a protein sequence. The refinement of domain annotations affected the

consistency of annotations for domain models of all lengths, but did not completely resolve the issue of incorrectly missing annotations for short domains. Therefore, some of the inferred changes in domain architectures are still likely not to be true evolutionary changes, but rather related to imperfect domain assignments.

In conclusion, refinement of domain assignments improved the quality of domain annotations and allowed me to be more confident when comparing domain architectures of proteins in the same phylogenetic tree. Additionally, this showed that phylogenetic information can in general be used as a tool for improving domain annotations in proteins.

Table 2.1 Increase of TreeFam proteins coverage. Sequence and residue coverege of proteins in the TreeFam database, before and after the refinement of domain assignments, is shown.

Measure	Before the refinement	After the refinement
Sequence coverage	84%	88%
Residue coverage	42%	46%

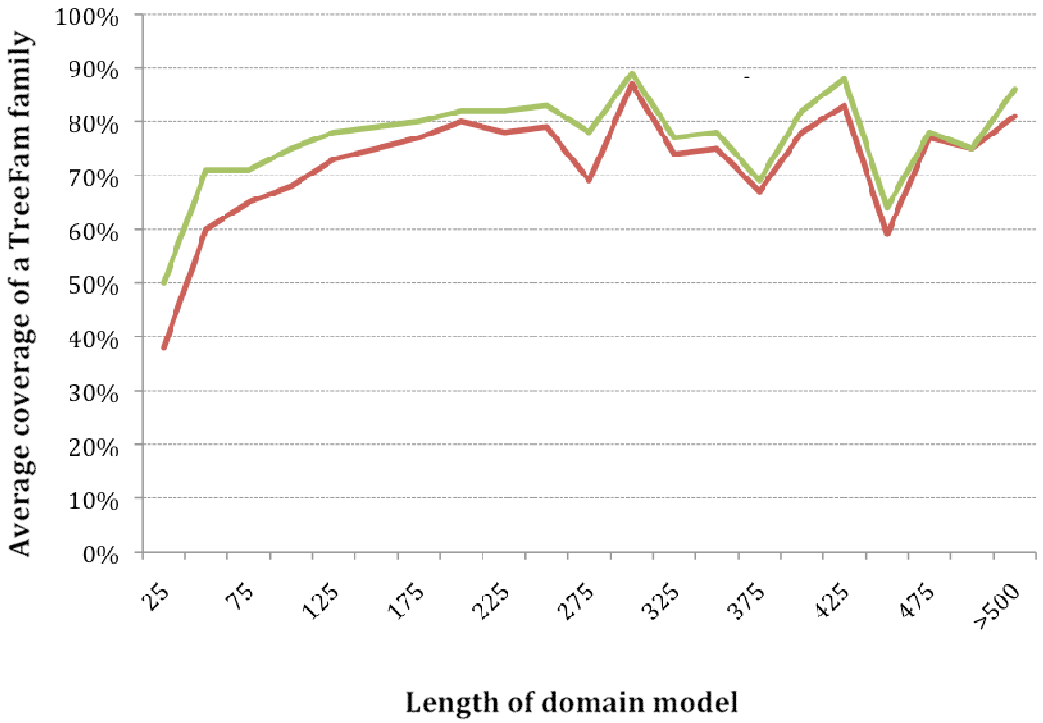

Figure 2.2: Average coverage of TreeFam gene families with Pfam domains of different lengths. Consistency of domain annotations between the members of the same TreeFam-A family represents the quality of domain annotations. Model lengths are grouped in bin categories of 25 amino acids, and all domains with model lengths longer than 500 amino acids are grouped together. The red line is showing the average coverage of TreeFam families with initial domain assignments and the green line after the refinement of domain assignments.

2.3.2 Single copy domains are predominantly gained and lost at protein termini

Previous comparisons of homologous proteins reported that changes in protein domain architectures preferentially occur at protein termini (Bjorklund et al., 2005; Weiner et al., 2006). I investigated here whether the same bias could be observed by directly following the evolution of an individual protein. This approach, using a protein's phylogenetic tree for the study of domain architecture evolution, has several advantages. Firstly, it is possible to infer the domain composition of an ancestral protein and hence the directionality of changes, i.e. distinguish domain gains from losses. Next, it is also possible to tell whether a change in the architecture occurred after gene duplication or after organism speciation. Finally, if the same change occurred multiple times, it is possible to map these events onto the tree and count the exact number of times when a certain domain architecture was formed. A comparison of homologous proteins that differ in domain composition, without using the associated phylogenetic information, cannot detect the cases of convergent evolution. To identify domain gain and loss events, I applied the maximum parsimony algorithm. The assumption here is that domain gains and losses are equally likely to occur. Additionally, I took into account only those changes that were supported with two or more descendant proteins – i.e. changes that were reported for internal nodes in the trees. This was necessary in order to avoid the effect of erroneous gene annotations - which were most likely to affect individual proteins.

First, I investigated the trends in gains and losses of domains that are not present as repeats in proteins; I call these domains 'single copy domains' here. The study of changes in the number of domains in repeats is described in Section 2.3.3. For each node in a tree where the inferred domain architecture of descendants differed from the inferred domain composition of an ancestral protein, I noted the position in the domain architecture where the change occurred. I separately studied changes that occurred after gene duplication from those that followed organism speciation. This allowed me to investigate if there were any differences - either due to the mechanisms or selective forces – that acted on proteins after these two types of evolutionary events. For each position,

N-, C-terminus, or middle, I also calculated the expected number of changes based on the expectation that a change is equally likely to occur anywhere in domain architecture.

I observed a strong positional bias for the changes to occur at the protein termini, rather than in the middle of proteins (Figure 2.3); the observed distribution of the number of changes at each position was significantly different from the expected one for all categories of events (P-value was always $< 2.2 \times 10^{-16}$, Chi-square test, Table 2.2; 2.2×10^{-16} is the smallest value in R for this test). This lent further support to reports from the previous studies (Bjorklund et al., 2005; Weiner et al., 2006). Interestingly, the bias was present both for the changes classified as domain gains and those classified as losses. Similarly, the same pattern was present irrespective of whether the change occurred after gene duplication or after speciation (Figure 2.3). Different molecular mechanisms can underlie gains and losses of domains (Babushok et al., 2007). Hence, it is interesting to observe that the same positional bias – for the changes to occur at the termini - exists when a domain is inserted into an ancestral protein and when it is deleted from it. On the other side, the same mechanisms for domain rearrangements should be available in the cell after gene duplication and speciation events. Hence, the observed similar patterns of positional bias for the changes following these two types of evolutionary events were in agreement with expectations.

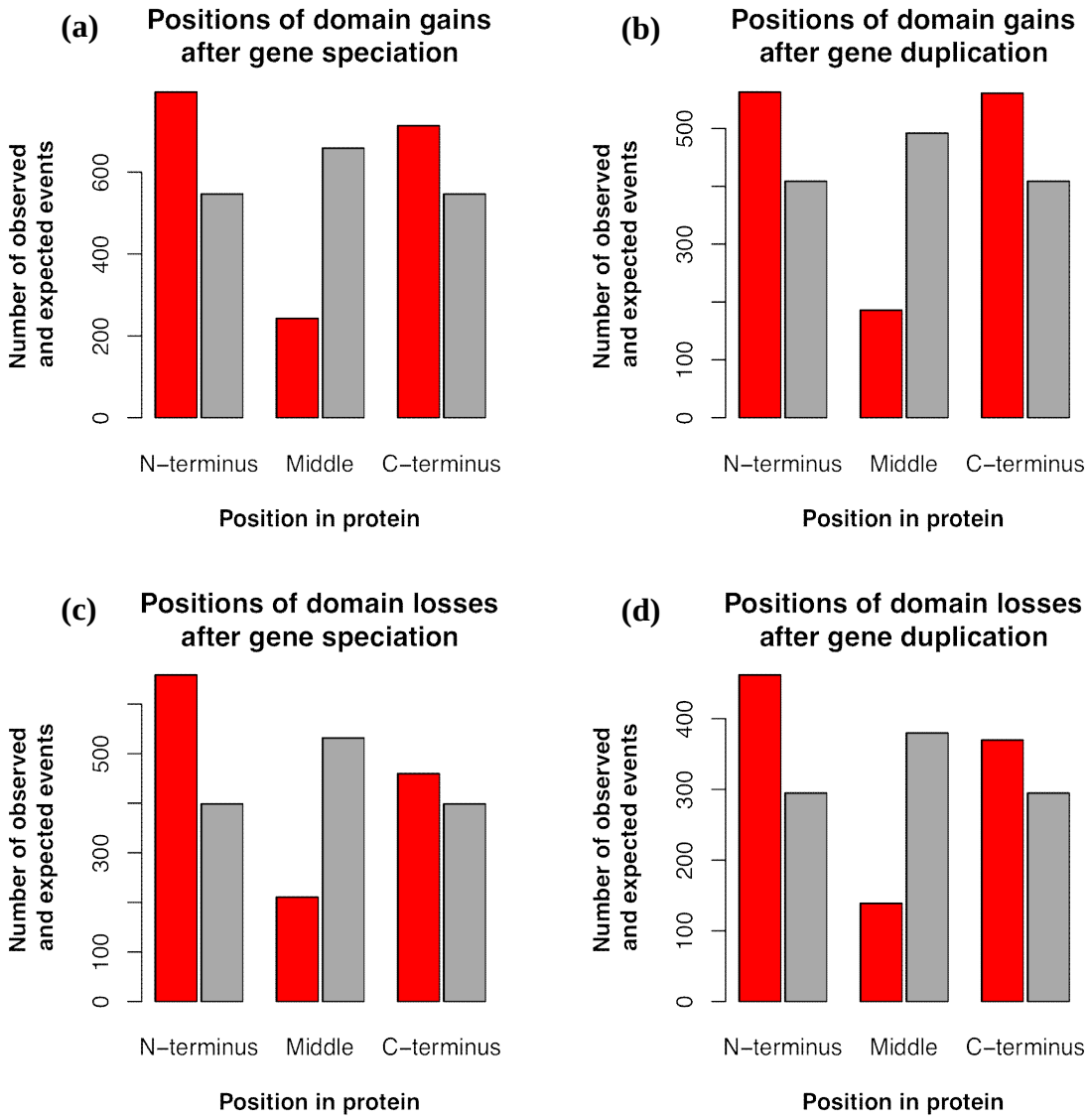

Figure 2.3: Positions of changes in proteins. Positions in proteins where gains (a and b) and losses (c and d) of single copy domains have been observed after gene speciation (a and c) and duplication (b and d) are shown. Observed and expected numbers of events are presented as red and grey columns, respectively. Observed numbers of events were obtained by applying maximum parsimony algorithm. Expected numbers of gains and losses were calculated based on the representation of ancestral proteins as strings of domains and an assumption that it is equally likely to observe a gain or loss of a domain at any position in the string. The presented data include single copy domains only. The bias for the changes to occur at the termini is evident in all categories of events.

Table 2.2: Statistical significance of the observed bias in positions of changes. Observed and expected numbers of changes at each position is indicated. P-value for the comparison between the two is obtained with a Chi-square test.

Evolutionary event	Change in domain architecture	Position of change	Number of observed events	Number of expected events	P-value
Speciation	Domain gain	N-terminus	796	547	$P<2.2 \times 10^{-16}$
		Middle	243	659	
		C-terminus	714	547	
	Domain loss	N-terminus	659	399	$P<2.2 \times 10^{-16}$
		Middle	211	532	
		C-terminus	460	399	
Gene duplication	Domain gain	N-terminus	563	409	$P<2.2 \times 10^{-16}$
		Middle	186	492	
		C-terminus	561	409	
	Domain loss	N-terminus	462	295	$P<2.2 \times 10^{-16}$
		Middle	139	380	
		C-terminus	370	295	

2.3.3 Gains and losses of domains in repeats

Changes in the number of domains in a repeat, i.e. of domains that exist as adjacent copies in a protein, can be caused by different molecular mechanisms compared to gains and losses of single copy domains (Bjorklund et al., 2006). For example, gains can occur through duplication of a region that encodes a domain and losses through deletion of a repetitive region during replication of genetic material in germ cells (Bjorklund et al., 2006). Similarly, evolutionary selection is likely to differently affect protein's evolution after the change in the number of domains in a repeat and after the gain or loss of a single copy protein domain. For example, duplication of an already existing domain can result in functional redundancy, but insertion of a new domain can cause a conflict in protein function. Similarly, repeating domains are often short – such as the leucine rich repeat family or C2H2 zinc fingers (Bjorklund et al., 2006) and hence, a change in the number of these domains is less likely to cause a larger structural disturbance. Therefore, the evolution of domain repeats has previously been studied separately (Bjorklund et al., 2006), and I also addressed it as a separate problem in this work.

The evolution of domain repeats is more complex to study than the changes in the overall domain composition of a protein. Firstly, many domains that occur in repeats are short and therefore are more likely to be omitted in the annotation process (see section 2.3.1). As a result of this, one needs to be more careful when interpreting the inferred changes. Secondly, analysis of the evolutionary trends is not as direct as in the case of domains that exist in one copy only. For instance, when a domain is deleted from a repeat - just by looking at the domain architectures - it is not always possible to say which domain from an ancestral protein is missing (Figure 2.1). Similarly, when a new domain is added to a domain repeat, it is not always possible to distinguish this domain from the domains that were present in the ancestral protein (Figure 2.1). I took this into account when assigning positions of changes, and treated each possible event as equally likely. As a consequence of this, it was more difficult to detect

trends that defined evolution of domain repeats than those that directed gains and losses of individual domains.

The analysis of positions at which changes in the number of domain repeats were inferred did not reveal as strong a bias for the protein termini as was observed for gains and losses of single copy domains (Figure 2.4). In strong contrast with the pattern for single copy domains, in one instance – for domain gains after gene duplications – the number of observed events at the N-terminus was lower than expected (Figure 2.4 b). However, divergence from the expected distribution, which was calculated from the assumption that all positions were equally likely, was still statistically significant (Table 2.3). Bjorklund *et al.* previously reported that the gain of new domains in a repeat frequently occurs through duplication of internal domains (Bjorklund et al., 2006). Therefore, it was expected that the distribution of positions of domain gains and losses would differ from the one for single copy domains. However, the bias for the termini is still present here. This implies that a combination of molecular mechanisms and evolutionary forces that influence both single copy domains and domain repeats, together with the ones specific for domains in repeats, could be at play here. However, it is important to note that averaging over all possible events, that were able to explain the observed changes, possibly camouflaged less strong trends in the evolution of domain repeats.

Again, a distribution of the positions of changes was similar both for the inferred domain gains and losses, and also between the changes that were observed after gene duplication and organism speciation events (Figure 2.4). This shows that when a domain is gained or lost from a protein, the strongest factor that influences positional preference of this event is the fact whether a domain is a part of a repeat or whether it exists as a single copy in a protein. In the case of a single copy domain there will be a very strong preference for the change not to occur in the middle of a protein. If a domain is in a repeat, this pressure will be less strong. The pressure for positional preference seems to be less dependent on whether the change in the architecture is a domain gain or loss, or whether the change occurred after gene duplication or after speciation.

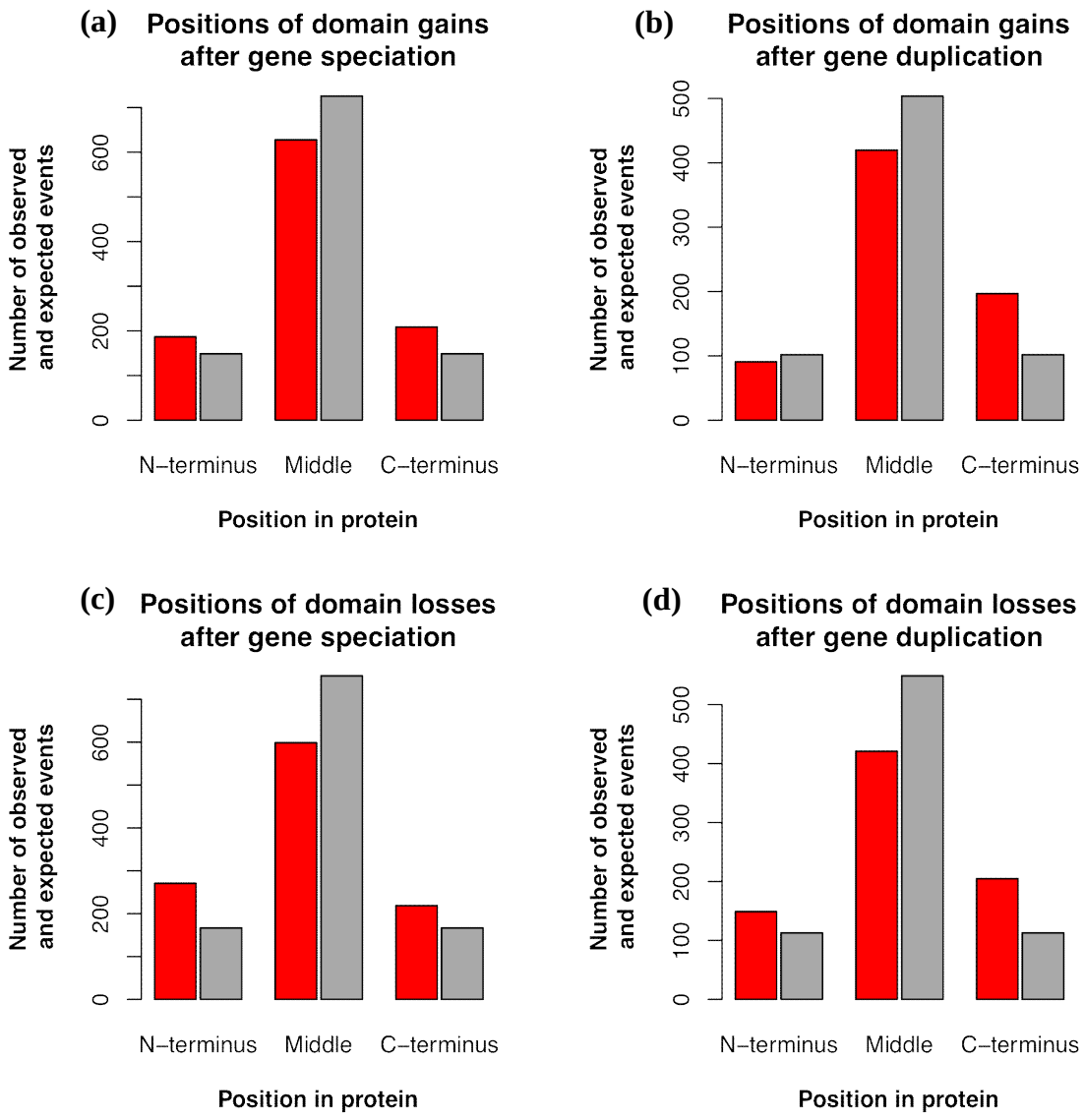

Figure 2.4: Positions of gains and losses of domains in repeats. Positions in proteins where gains (a and b) and losses (c and d) of domains in repeats have been observed after gene speciation (a and c) and duplication (b and d). Observed and expected numbers of events are presented as red and grey columns, respectively. Observed numbers of events were obtained by applying the maximum parsimony algorithm. When a position of a change was ambiguous all possible scenarios were taken into account and the number of changes was weighted with the probability of each event. Expected numbers of gains and losses were calculated based on the representation of ancestral proteins as strings of domains and an assumption that it is equally likely to observe a gain or loss of a domain at any position in the string. There is still bias for the changes to occur at protein termini, but this bias is not as strong as it is for single copy domains.

Table 2.3: Comparison between distributions of observed and expected number of domain gains and losses at each position in a protein for the changes in the number of domains in repeats. Observed and expected number of changes at each position is indicated. P-value for the comparison between the two distributions is obtained with Chi-square test.

Evolutionary event	Change in domain architecture	Position of change	Number of observed events	Number of expected events	P-value
Speciation	Domain gain	N-terminus	187	149	$P<3.5 \times 10^{-11}$
		Middle	628	726	
		C-terminus	209	149	
	Domain loss	N-terminus	271	167	$P<2.2 \times 10^{-16}$
		Middle	599	755	
		C-terminus	219	167	
Gene duplication	Domain gain	N-terminus	91	102	$P<2.2 \times 10^{-16}$
		Middle	420	504	
		C-terminus	197	102	
	Domain loss	N-terminus	149	113	$P<2.2 \times 10^{-16}$
		Middle	421	549	
		C-terminus	205	113	

2.3.4 Changes in domain architectures preferentially occur after gene duplications

The evolution of domain architectures does not necessarily need to follow the same pattern after gene duplication and after organism speciation. This is why I separately investigated domain gains and losses that occurred after these evolutionary events. As discussed in sections 2.3.2 and 2.3.3, there was no significant difference in the positional preference between the changes that followed gene duplications and those that followed organism speciation. However, the total number of gene duplication events, or duplication nodes in the TreeFam trees, is smaller then the total number of speciation events/nodes, and the number of observed changes was higher after gene duplications (Figures 2.3 and 2.4). Therefore, I compared the frequency of changes after gene duplication and speciation events (Table 2.4). On average, change in the overall domain composition, i.e. gain or loss of a single copy domain, is observed after 87 speciation events but almost twice as frequently after gene duplications; on average once in 43 gene duplication events. Similarly, a change in the number of domains in a repeat occurs on average after 128 speciation events, in comparison to after on average 67 gene duplication events; again almost two times more frequently after gene duplications.

As an additional test, I compared the branch lengths in TreeFam trees before gene duplication and speciation events for which the changes were inferred. This again showed that the average branch length, or the average time span, before a domain was gained or lost from a protein was about twice as long for speciation compared to gene duplication, irrespective of whether the domain existed as a single copy domain in a protein or was a part of domain repeat (Table 2.4). The branch lengths are based on the similarity of proteins and hence are influenced by the presence or absence of a protein domain. Therefore, this only gives an indication of the evolutionary time that passed before a domain was gained or lost. Nonetheless, both means for calculating the frequency of changes in domain architectures showed that there was a bias for the changes to preferentially occur after gene duplications. Table 2.4 shows the total number of internal nodes and a sum of branch lengths in all TreeFam trees that I used in

calculations. The total number of inferred changes of domain architecture for gene duplication and speciation events was calculated from the data in Tables 2.2 and 2.3.

Table 2.4: Changes in domain architecture occur more frequently after gene duplications then after organism speciation. Frequency of the change is stated as an average number of events for which the change is observed and as an average branch length before the change is observed. Calculations include all TreeFam trees.

Domain affected	Evolutionary event	Number of nodes in TreeFam trees	Total branch length before all events of this type	Average number of events for which the change is observed	Average branch length before the change is observed
Single copy domain	Speciation	269478	34342.29	87	11.14
	Gene duplication	99106	13526.49	43	5.93
Domain in repeat	Speciation	269478	34342.29	128	16.25
	Gene duplication	99106	13526.49	67	9.12

2.3.5 Effect of domain gains on the evolution of protein function

Gains and losses of protein domains are likely to strongly influence the overall protein function. If having a protein with new domain architecture is disadvantageous for the organism, the protein will probably be removed from the population. Therefore, domains that are observed as frequently gained have likely conferred functional advantage to proteins, which they were inserted in. The most often gained domains from this study are listed in Table 2.5. The table includes only domains gained on the internal nodes of the TreeFam trees. All these domains belong to one of the following functional categories: extracellular processes, regulation through signal transduction or regulation through DNA binding. Hence, those domains that act as modifiers of the overall function, rather than domains with a specific function, are more likely to combine with other protein domains and be useful in different cellular contexts. Domains with extracellular function are the EGF (epidermal growth factor) superfamily, the immunoglobulin domain and the CUB (complement protein subcomponents C1r/C1s, urchin embryonic growth factor and bone morphogenetic protein 1) domain, and those that act as signal transducers are zinc finger (C2H2 type), leucine-rich repeat, SH3 (Src homology 3) domain, the PH (pleckstrin homology) domain and RING (really interesting new gene)-finger superfamily.

Additionally, functional compatibility between a gained domain and domains present in the ancestral protein also decides on whether the new protein will be useful to a cell. I used a method for comparing GO terms (Schlicker et al., 2006), which were projected to Pfam domains, to estimate functional similarity between gained and ancestral domains. The score for the similarity measure, *funSim*, that I used here ranges from 0 to 1 with a score close to 1 corresponding to GO terms with highly similar function and those below 0.3 to GO terms that are not functionally related. I found that only 454 internal domain gain events were applicable for this analysis, meaning they had both gained and ancestral domains annotated with GO terms and *funSim* scores available for the annotated terms. Interestingly, only 18% of the gained domains were not functionally similar (*funSim* < 0.4) to any domain in the ancestral protein (81 out of 454 events). The other gained domains were reported to be

functionally related to at least one domain in the ancestral protein, and 39% of the gained domains (176 out of 454 events) highly similar to a domain in the ancestral sequence (*funSim* > 0.8). This implies that domain gain usually does not radically change the protein function, but only adapts it to new contexts.

Table 2.5: Most frequently gained domains in animal phylogenetic trees. Pfam IDs, domain/clan descriptions and associated functional categories of domains that are most frequently gained in all TreeFam trees are listed in the table.

Number of observed gains	Pfam ID	Domain description	Functional category
115	CL0001	EGF superfamily	Extracellular processes
87	CL0159	Ig-like fold superfamliy	Extracellular processes
85	PF00096	Zinc finger, C2H2 type	Regulation: DNA-binding
76	CL0011	Immunoglobulun superfamily	Extra cellular processes
66	CL0164	CUB domain	Extracellular processes
65	CL0022	Leucine rich repeat	Signal transduction/ Extra cellular processes
60	CL0266	PH domain-like superfamily	Regulation: Signal transduction
56	CL0010	Src homology-3-domain	Regulation: Signal transduction

2.3.6 Estimate of domain gain and loss events strongly depends on the input parameters

Domain gain and loss events that I discussed in the sections 2.3.2 – 2.3.5 are inferred from the assumption that gains and losses are equally likely and that differences in domain architectures of related genes can be explained with as few changes as possible. However, there is no general consensus on what the relative frequencies of these events are. Different studies have used different values for the frequencies of domain gain and loss events and applied maximum, weighted or Dollo parsimony to infer changes in domain architectures (Basu et al., 2008; Fong et al., 2007; Itoh et al., 2007). In this section, I investigate how much the estimate of the likelihood of these events influences whether the present domain architectures are explained by ancestral gain or loss events. For this, I applied a weighted parsimony algorithm. By changing the costs, or weights, for domain gain and loss, I was able to change the assumptions about the frequency of these events. I found that the total number of inferred gain or loss events was strongly influenced by the initial estimates of their frequency (Figure 2.5). Again, to avoid the effect of erroneous gene annotations, I included in the analysis only changes observed on the internal nodes in the trees. The ratio of reported gains over losses (Figure 2.5b) - and the ratio of reported losses over gains (Figure 2.5a) - exponentially increased as the assumed probability for the ratio of events linearly increased. Figure 2.5c shows a logarithmic representation of these values. The expected, or assumed, ratio of observed changes is indicated by a red line and the observed, i.e. inferred, one by blue dots. The assumed probabilities of gain and loss events determined the observed ratios to a higher degree then expected.

These calculations showed that inferred evolutionary scenarios are strongly influenced with their initially estimated likelihoods. When the input parameters for the cost of domain gain and loss are equal, the observed number of domain gains and losses is also about the same. This is the scenario, which is applied in the maximum parsimony algorithm. Hence, this stresses that one should be careful when interpreting observed gains and losses in these kinds of studies. Furthermore, it shows that in order to obtain a confident set of gain or

loss events one needs to be very careful about the algorithm and parameters used.

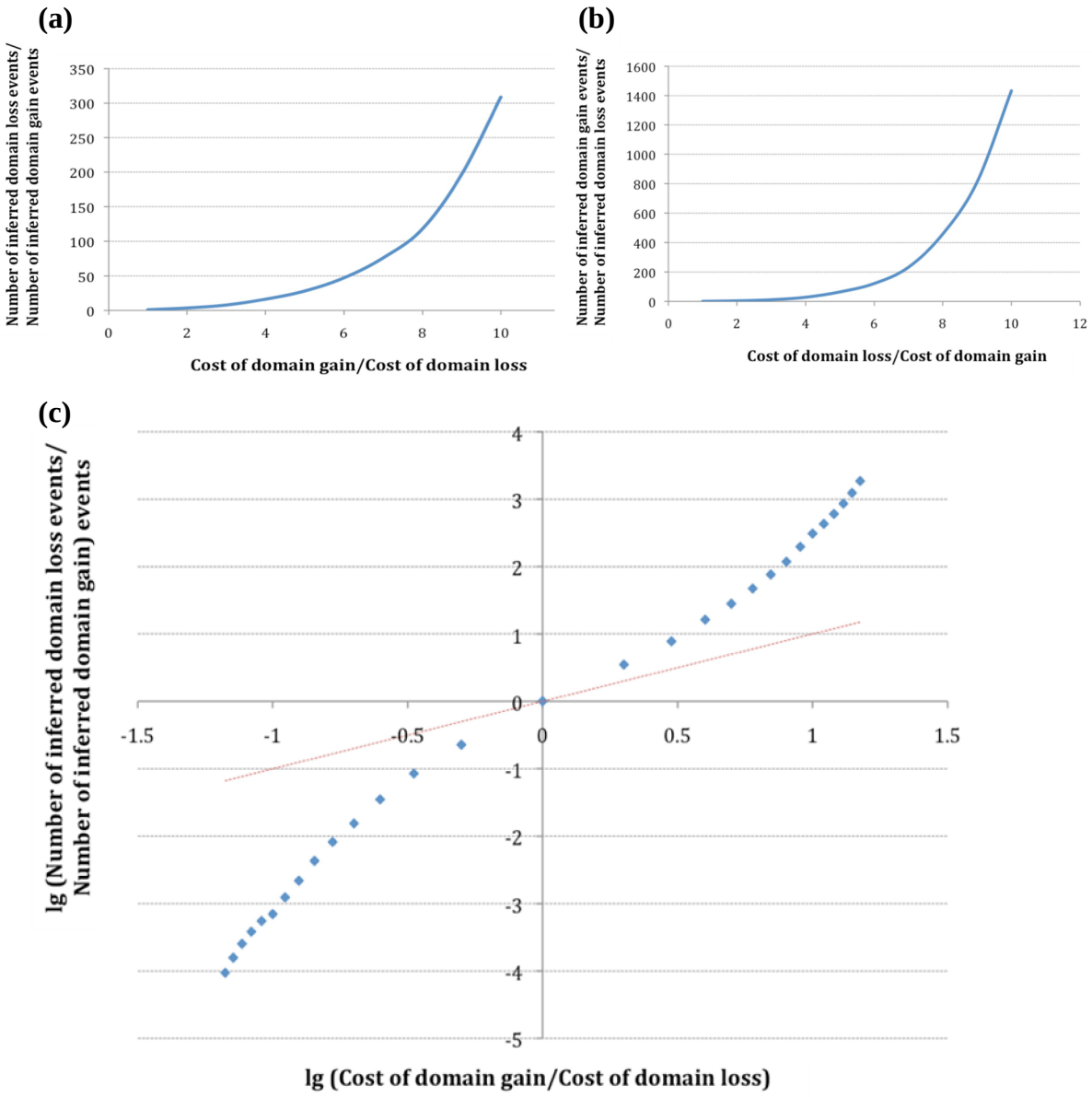

Figure 2.5: The ratio of inferred domain gain and loss events strongly depends on the assumed cost of these events. (a) The ratio of inferred domain loss and gain events exponentionally depends on the ratio of increasing assumed cost for domain gain and loss event. The higher the cost of an event, the smaller is the likelyhood of observing the event. (b) Similarly to (a), increasing the cost of domain loss results in an exponentional increase of the inferred ratio of domain gain and loss events. (c) Logarithmic representation of the data on graphs (a) and (b). The red dotted line reprents the logarithm of the expected ratio of domain loss and gain events as assumed by the weights for these events. Blue data points show the log values of the inferred ratio of these events. The inferrend ration shows a strong divergence from the expected one.

2.4 Discussion

2.4.1 Confidence in the comparison of domain architectures

The aim of the research described in this chapter was to investigate the general trends in the evolution of protein domain architectures. For annotation of proteins with domains and families, I used Pfam-A protein families. Pfam-A release 22, that I used here, had nearly 10,000 protein families. This ensured much better coverage of proteins with domain assignments than it would have been possible if, for example, structural domain annotations had been used. Additionally, Pfam-A domains are of very good quality and provide literature references for the domains. Hence, after domain gain or loss event, it is often possible to analyse consequences of the event on the overall protein function. Inclusion of Pfam-B families in the study would have further increased the protein coverage with domain assignments and, because of that; a greater number of changes in protein domain architectures would have been detected. However, Pfam-B families are in general of lower quality than Pfam-A families, and those composed of low complexity regions may not even reflect true evolutionary relationships. Therefore, to increase the confidence of observed domain gains and losses, I included only Pfam-A families in the study.

Apart from reflecting true changes in domain architectures, apparent changes of domain composition can also be a result of incomplete domain annotations or erroneous gene assignments. To overcome these issues, I adjusted the procedure for identifying domain gains and losses. When the inconsistency of domain assignments in a TreeFam family was not justified with significant differences on the protein sequence level, I added domains to the family members that initially lacked them. Additionally, I excluded from the analysis the cases where changes in protein domain composition were not supported by at least two descendant proteins. The main reason for doing this was to avoid the effects of incomplete gene annotations. Both refinement steps were done in order to obtain a set of inferred domain gain and loss events enriched in the events that describe real changes of domain architectures.

Alternatively, apparent differences in domain composition can also assist gene and domain annotation methods. For example, when domain assignments of a single protein in a phylogenetic tree differ from the ones of its homologues, this might be also because not all of the exons are predicted for this gene. In particular, genes from the genomes with lower quality annotations, which lack domain assignments, could be the candidates for an assessment and refinement of their gene boundaries. Additionally, as described in the section 2.3.1, phylogenetic trees can be used as a tool to guide the refinement of imperfect initial domain annotations. The approach that I applied here is similar to previously described context analyses, in a sense that in order to improve protein annotations, it uses the information about domains present in related proteins. Additionally, this approach, for the first time, utilizes phylogenetic relations among proteins as an incentive for examining similarity in the protein regions with inconsistent domain assignments.

The increase of TreeFam coverage that this resulted in (Table 2.1) shows that this approach can in general be used to assist protein annotation.

2.4.2 Molecular mechanisms and evolutionary selection shape the evolution of domain architectures

I have investigated here several aspects of protein domain architecture evolution, including positions of changes in proteins, their frequency after gene duplication and speciation events, and function of the most frequently gained domains. Characteristics of the present domain architectures reflect the interplay of molecular mechanisms and evolutionary selection that shaped their evolution. One of the crucial observations from previous work on protein evolution, which came from the comparison of homologous proteins, was that changes in domain architecture preferentially occur at the N- and C- termini (Bjorklund et al., 2005; Weiner et al., 2006). Weiner *et al.* described this observation with the fact that the dominating mechanisms that caused the changes are those that acted at protein termini. Hence, they proposed that the evolution of novel proteins was mainly defined with gene fusion and fission events and in particular, insertions of new start and stop codons. Here, by using

gene phylogenies, I was able to distinguish between inferred domain gain and loss events. Interestingly, even though there are molecular mechanisms that result only in domain gains or only in domain losses, both categories of events showed strong bias towards protein termini, particularly in the case of gains and losses of single copy domains. Therefore, the observed distribution of changes is better explained with the interplay of both: mechanisms that acted to add or remove domains at the protein termini, as well as evolutionary selection that disfavoured domain gains and losses within a protein (Figure 2.6a). Protein termini are normally charged, flexible and found at protein surface (Figure 2.6b), so it is easy to imagine that additions or deletions of domains there are less likely to disrupt the rest of the structure, especially if the concerned domains are independent structural units. On the other hand, connector regions between domains direct the contact and interaction of domains they link together. Hence, even if those regions themselves are unstructured and do not have a functional role; it is still more likely that changes there will disrupt the rest of the structure. Because of this, evolutionary selection is likely to strongly favour changes at the termini over the changes in the middle of proteins. Since I compared here only the overall domain architectures, I could not directly infer the positions of insertion and deletion of domains in repeats. Additionally, changes in the number of domains in repeats are particularly difficult to study in general. Many domains in repeats are short and therefore their assignments to proteins are often not of high confidence (Figure 2.2). Therefore, the inferred gains and losses of repeated domains in this study are of lower confidence than those of single copy domains. To overcome the issue of omitted domain assignments, one possibility is to lower the threshold for assignment of domains in repeats (Bjorklund et al., 2005). However, this again increases the chance of false positive domain annotations.

The observed trends in the evolution of domain repeats imply that the positional bias is not as strong as it is for insertions and deletions of single copy domains. It is possible that additional mechanisms, which do not have a positional preference, such as duplication and deletion of sequence repeats after misalignment of homologous alleles (Bjorklund et al., 2006), play an important role in their evolution and hence influence the overall pattern of changes.

Nonetheless, even domain repeats with changes at the termini possibly have a smaller effect on the structural stability and hence a higher chance to go through evolutionary selection. The combination of acting mechanisms and evolutionary selection drives both changes in single copy domains and changes in the number of domains in repeats.

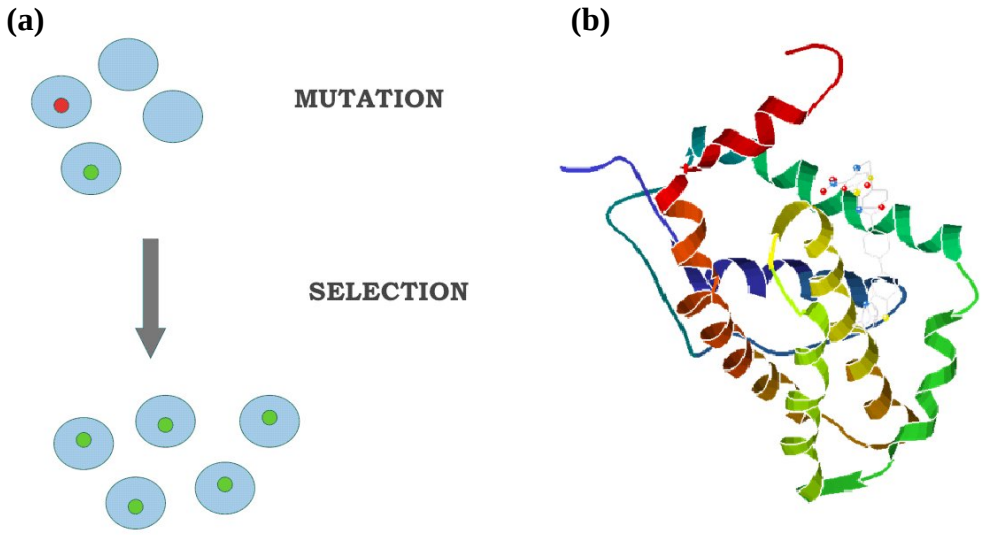

Figure 2.6: The evolution of domain architectures is determined by molecular mechanisms that cause the changes as well as subsequent selection. (a) Different molecular mechanisms can cause changes in domain architecture, but only some of the created architectures survive the subsequent evolutionary selection. Red and green dots represent mutated proteins in different individuals. After evolutionary selection only a mutation shown as a green dot became fixed in a population. (b) Protein's structural stability can have a strong influence on the selection of novel domain architecture. The charged termini are usually found on the protein's surface and changes at the surface are less likely to severely disrupt the overall structure. This is illustrated with a structure of the anti-apoptotic Bcl-2 protein.

TreeFam phylogenies distinguish between gene duplication and organism speciation events. Comparison of the positions of changes, which followed these two types of evolutionary events, did not show a difference in trends. This implies that the same basic mechanisms and evolutionary forces influenced emergence of new domain architectures and drove evolution of an individual protein both after gene duplication and after speciation. However, the frequency with which the changes are observed is nearly two fold greater after gene duplications (Table 2.4). This suggests that an important difference between the

two types of events is played by evolutionary selection, which is more permissive towards changes in proteins when the original gene exists in two copies and the introduced changes do not imply complete loss of the ancestral function (Zhang, 2003).

Domains that were most frequently gained during animal gene evolution either have a role in extracellular processes or in cell regulation - such as signal transduction or DNA binding (Table 2.5). Interestingly, Vogel and Chothia (Vogel and Chothia, 2006) reported previously that the number of genes in an organism with these same domains (apart from the leucine-rich repeat protein family) is in a strong correlation with organism complexity. In accordance with this, they have suggested that these domains were responsible for the emergence of new complex traits in metazoans. Vogel and Chothia (Vogel and Chothia, 2006) have assigned the expansion of these domains primarily to duplications of the genes that already contained them. However, this study implies that insertion of these domains into genes that have not previously coded for them has also contributed to their expansion. Hence, not only duplication of these domains, but their combination with other domains could have played a role in the evolution of novel, animal specific, traits. Additionally, when functional annotation of both ancestral and gained domains was available, the study showed that in the majority of the cases the gained domain was of the similar function as the ancestral domains. This is in agreement with previous studies that showed that gene fusion usually occurs between genes of similar function (Yanai et al., 2001) and once again underlies the role of evolutionary selection, which over time eliminates from the population domain combinations that are not likely to confer an advantage to the organism.

In conclusion, protein evolution is evident at different scales of events. On the small scale, single amino acids are mutated, and, on the large scale, whole domains are lost or gained in the protein. The observed changes are primarily defined with the molecular mechanisms that cause the mutations. However, selective constraints imposed by the necessity for structural stability and for the functional protein product also play a crucial role in protein evolution. Of course, a protein's function and evolution is defined not only by its sequence, but also by its genomic position, expression pattern, and partners in its interaction network

and a systematic approach is needed to fully understand the evolutionary path of an individual protein (Pal et al., 2006).

2.4.3 Set of confident domain gain or loss events

Novel domain architectures are the result of a joint action of mechanisms that created them and subsequent evolutionary selection. Hence, the observation that changes preferentially occur at the termini also implies that molecular mechanisms that act at protein termini are the ones that play the most important role in protein evolution. However, to draw concrete conclusions about the relative contributions of different mechanisms it is important to firstly obtain a set of confident domain gain or loss events. In the section 2.3.6, I have showed that inference of domain gains and losses is strongly influenced by the applied algorithm and assumed probability of these events. Therefore, even though inference of domain gains and losses by the maximum parsimony algorithm gives an indication of general trends in the evolution of protein domain composition, it does not provide a high enough quality set of events for the further investigation of the causative mechanisms. In Chapter 3, I am discussing the approach that I applied to obtain such a confident set of domain gains and the analyses I performed to investigate evidence for the action of each possible mechanism. I focus the study on domain gains and the evolution of more complex domain architectures. As indicated also here by the character of the most frequently gained domains (Table 2.5), the addition of novel domains to proteins likely played a crucial role in the evolution of complex animal traits. However, domain losses also change the function of the resulting protein products and protein evolution through domain loss could be an important mechanism for subfunctionalization of proteins.

2.5 Bibliography

Babushok, D.V., Ostertag, E.M., and Kazazian, H.H., Jr. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci *64*, 542-554.

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.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic acids research *30*, 276-280.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. (2000). The Pfam protein families database. Nucleic acids research *28*, 263-266.

Beaussart, F., Weiner, J., 3rd, and Bornberg-Bauer, E. (2007). Automated Improvement of Domain ANnotations using context analysis of domain arrangements (AIDAN). Bioinformatics (Oxford, England) *23*, 1834-1836.

Bjorklund, A.K., Ekman, D., and Elofsson, A. (2006). Expansion of protein domain repeats. PLoS computational biology *2*, e114.

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.

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. Nucleic acids research *33*, D212-215.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. Science (New York, NY *300*, 1701-1703.

Coin, L., Bateman, A., and Durbin, R. (2003). Enhanced protein domain discovery by using language modeling techniques from speech recognition. Proceedings of the National Academy of Sciences of the United States of America *100*, 4516-4520.

Coin, L., Bateman, A., and Durbin, R. (2004). Enhanced protein domain discovery using taxonomy. BMC bioinformatics *5*, 56.

Das, S., and Smith, T.F. (2000). Identifying nature's protein Lego set. Advances in protein chemistry *54*, 159-183.

Ekman, D., Bjorklund, A.K., Frey-Skott, J., and Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. Journal of molecular biology *348*, 231-243.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86-90.

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.

Gough, J. (2005). Convergent evolution of domain architectures (is rare). Bioinformatics (Oxford, England) *21*, 1464-1471.

Holm, L., and Sander, C. (1994). Parser for protein folding units. Proteins *19*, 256-268.

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.

Kummerfeld, S.K., and Teichmann, S.A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet *21*, 25-30.

Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., *et al.* (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic acids research *34*, D572-580.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science (New York, NY *285*, 751-753.

Moore, A.D., Bjorklund, A.K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. Trends in biochemical sciences *33*, 444-451.

Oliver, P.L., Goodstadt, L., Bayes, J.J., Birtle, Z., Roach, K.C., Phadnis, N., Beatson, S.A., Lunter, G., Malik, H.S., and Ponting, C.P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics *5*, e1000753.

Pal, C., Papp, B., and Lercher, M.J. (2006). An integrated view of protein evolution. Nature reviews *7*, 337-348.

Pasek, S., Risler, J.L., and Brezellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics (Oxford, England) *22*, 1418-1423.

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling--a review. Gene *238*, 103-114.

Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics *7*, 302.

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. The FEBS journal *273*, 2037-2047.

Yanai, I., Derti, A., and DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proceedings of the National Academy of Sciences of the United States of America *98*, 7940-7945.

Zhang, J. (2003). Evolution by gene duplication: an update. TRENDS in Ecology and Evolution *18*, 292-298.