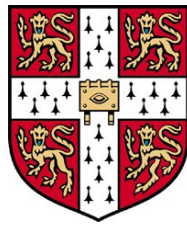


# Mechanisms of change in protein architecture



Marija Buljan  
Trinity College  
University of Cambridge

A dissertation submitted for the degree of  
Doctor of Philosophy  
September 2010

“Well begun is half done.”

Aristotle

## Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between May 2006 and September 2010. This dissertation is the result of my own work. No part of this dissertation or anything substantially the same has been or is being submitted for any qualification at any other university.

## Summary

Proteins are the basic building blocks and functional units in all living organisms. Moreover, differences between species can frequently be explained with differences in their protein complements. Importantly, proteins are often composed of segments, i.e. domains that have a certain level of evolutionary, structural and/or functional independence. The majority of proteins in nature contain two or more domains, and an individual domain can often occur in combinations with different domain partners.

In the first part of my thesis, I traced the history of animal gene families and the proteins these genes encode. By this means, I was able to infer events where changes in protein domain architectures took place. This showed that both insertions and deletions of single copy domains preferentially occur at protein termini, but also that changes are more likely to occur after gene duplication than organism speciation. Finally, domains that were most frequently gained were the ones that are related to an increase in organismal complexity, thus underlining the important role of domain shuffling in animal evolution.

In the second part of my thesis, I focused on a set of high confidence domain gain events and investigated the evidence for molecular mechanisms that caused these domain gains. In agreement with observations from the first part - that changes preferentially occur at the termini - I have found that the strongest contribution to gains of novel domains in proteins comes from gene fusion through the joining of exons from adjacent genes into a novel gene unit. Two other mechanisms that have been suggested to play a major role in the evolution of animal proteins, retroposition and middle insertions through intronic recombination, have a smaller role in comparison to gene fusions. Since the majority of these domain gains are again observed after gene duplication, this suggests a powerful mechanism for neofunctionalization after gene duplication.

Finally, in the last part of my thesis, I address a mechanism that increases the number and variety of proteins in an organism – alternative splicing. In particular, I investigate the functional consequences of tissue-specific alternative splicing events. I found that tissue-specific splicing tends to affect exons that encode protein regions without defined secondary or tertiary structure. Importantly, it is known that these disordered regions frequently play a role in protein interactions. In agreement with this, I observed significant enrichment of tissue-specifically encoded protein segments in disordered binding peptides and posttranslationally modified sites. A possible result of the finely regulated alternative splicing of these segments is a tissue-specific rewiring of protein network. In conclusion, both alternative splicing and domain shuffling can increase proteome diversity. However, a protein with a new function can often directly or indirectly shape the functions of other proteins in its environment.

## Acknowledgments

During my PhD, I was lucky to meet many people that I will remember as inspiring scientists and great persons. In these acknowledgments, I wish to say thanks to those individuals that most influenced the work that I deliver here. I am very grateful to my supervisor Alex Bateman for his support during my PhD, for teaching me how to approach science critically and with attention to details and for providing me a valuable feedback on everything that is written in this thesis. My thesis committee meetings were of immense value for my progress in the PhD. Madan Babu, Richard Durbin, Avril Coghlan and Manolis Dermitzakis, my thesis committee members, contributed with great ideas and most helpful criticism. I am also indebted to Avril for her help at the beginning of my PhD when I was starting with phylogenetic analyses. Madan gave valuable contribution to the work in Chapter 4, which is also a joint project with him, and I am very grateful to him for that. The Xfam group members were extremely helpful during my PhD, providing advice and support when I needed it. I am particularly grateful to Benjamin Schuster-Bockler, Cara Woodwark, Lars Barquist, Paul Gardner and Rob Finn. I also need to thank other students in my year, who were always available for discussions, gave valuable feedback so many times and provided excellent PhD environment. I need to specially thank there Matias Piipari and Leo Parts for their help. Finally, Neil Rawlings, Paul Gardner and Penny Coggill kindly proofread the thesis chapters and provided the most useful comments.

# Contents

1. Introduction	1
1.1. Characterization of functional elements in proteins.....	3
1.1.1. Protein domains.....	3
1.1.2. Disordered protein regions.....	7
1.1.3. Sites of posttranslational modification.....	11
1.2. Protein evolution.....	12
1.2.1. Domain shuffling.....	14
1.2.2. Mechanisms for formation of novel genes.....	18
1.2.3. Gene duplication and protein evolution.....	24
1.2.4. Evolutionarily related proteins.....	28
1.3. Protein isoforms of the same gene.....	29
1.4. Outline of the thesis.....	33
1.5. Bibliography.....	34
2. Evolution of multidomain proteins	45
2.1 Introduction.....	45
2.2 Methods.....	50
2.2.1 Analysis of TreeFam families.....	50
2.2.2 Assignment of domains to proteins with refinement.....	50
2.2.3 Domain gains and losses.....	51
2.3 Results.....	53
2.3.1 Phylogenetic trees can guide refinement of domain assignments.....	53
2.3.2 Single copy domains are predominantly gained and lost at protein termini.....	57
2.3.3 Gains and losses of domains in repeats.....	61
2.3.4 Changes in domain architectures preferentially occur after gene duplications.....	65

2.3.5	Effect of domain gains on the evolution of protein function.....	67
2.3.6	Estimate of domain gain and loss events strongly depends on the input parameters.....	69
2.4	Discussion.....	71
2.4.1	Confidence in the comparison of domain architectures.....	71
2.4.2	Molecular mechanisms and evolutionary selection shape the evolution of domain architectures.....	72
2.4.3	Set of confident domain gain or loss events.....	76
2.5	Bibliography.....	77
3.	Mechanisms of domain gain in animal proteins	80
3.1	Introduction.....	80
3.2	Methods.....	86
3.2.1	Assignment of domains to proteins with refinement.....	86
3.2.2	Exclusion of possible false domain gain calls.....	86
3.2.3	Parsing trees.....	87
3.2.4	Intron-exon structures of genes.....	91
3.2.5	Positions of gained domains.....	91
3.2.6	Genomic origin of the inserted domain.....	92
3.3	Results.....	94
3.3.1	Set of high confidence domain gain events .....	94
3.3.2	Characteristics of the high confidence domain gain events.....	95
3.3.3	Characteristics of the medium confidence domain gain events.....	97
3.3.4	Supporting evidence for the representative transcripts.....	99
3.3.5	Donor genes of the gained domains.....	101
3.3.6	Investigation of cellular mechanisms that caused domain gain events.....	102
3.3.6.1	Retroposition as a mechanism of domain gain.....	102
3.3.6.2	Joining of adjacent genes as a mechanism of domain gain.....	105
3.3.6.4	Insertion of exons into ancestral introns as a mechanism of domain gain.....	112
3.3.6.4	Exonisation of previously non-coding sequences as a mechanism of domain gain.....	113
3.3.7	Domain gains most frequently occur after gene duplications.....	115



3.3.8	Gained domains do not have their origin in the adjacent genes...	119
3.3.9	Domain gain events affect cellular regulatory networks.....	119
3.4	Discussion.....	123
3.4.1	Scope of the study.....	123
3.4.2	Approach for obtaining the set of confident domain gain events.....	124
3.4.3	Mechanisms of domain gain.....	125
3.4.4	Domain gains were assisted by recombination events.....	128
3.4.5	Different trends in domain gains in different lineages and at different time points during evolution.....	130
3.4.6	Functional implications of domain gain events.....	131
3.5	Bibliography.....	132
4.	Protein products of tissue-specific alternative splicing	138
4.1	Introduction.....	138
4.2	Methods.....	142
4.2.1	Sets of tissue-specific, cassette and constitutive exons.....	142
4.2.2	Enrichment of genes with specific function in the set of tissue-specific exons.....	143
4.2.3	Prediction of disordered protein residues.....	144
4.2.4	Prediction of functional residues.....	144
4.2.5	Conservation of exons in the three different datasets.....	145
4.2.6	Significance of observed trends.....	145
4.2.7	Comparison of MEK1 and MEK2 protein sequences.....	146
4.2.8	Enrichment of known disease genes in the set of tissue-specific exons.....	146
4.2.9	Disorder signatures in the protein products of the p73 gene.....	147
4.3	Results.....	147
4.3.1	Sets of exons with different expression profiles.....	147
4.3.2	Tissue-specific exons are enriched in disordered residues.....	149
4.3.3	Functional residues in disordered segments encoded by tissue-specific exons.....	151
4.3.4	Distribution of functional residues in the control sets of cassette and constitutive exons.....	154

4.3.5	Disordered residues encoded by tissue-specific exons are highly conserved.....	156
4.3.6	Genes with tissue-specifically regulated exons have an important function in organism development and survival.....	160
4.3.7	Alternative isoforms of the gene p73.....	165
4.3.8	Tissue-specific splicing and protein domains.....	167
4.4	Discussion.....	170
4.4.1	Evolution and function of alternative splicing.....	170
4.4.2	Unstructured functional residues direct isoform-specific networks.....	172
4.4.3	Examples for the role of disordered protein segments in signal transduction.....	175
4.4.4	Genes with tissue-specific isoforms and disease development...	178
4.5	Bibliography.....	180
5.	Concluding remarks	186
	Appendices	190
	Appendix A.....	191
	Appendix B.....	194
	Appendix C.....	204

# Chapter 1

## Introduction

Proteins are crucial functional elements of living organisms, involved in virtually every process within cells. Often, proteins with similar functions – which belong to the same or to different organisms - are evolutionary related. A well-described example for this is a family of oxygen-carrying globins in vertebrates. The major steps in the evolution of this family involved duplication of an ancestral oxygen-binding protein, divergence of the copies into myo- and haemoglobin, and another duplication and divergence of ancestral haemoglobin into alpha and beta subunits (H Lodish, 2000). These and other proteins from the same globin family are all involved in oxygen transport but have evolved subtle differences of function, which make them suited to specific roles in the physiology of oxygen transport. Since the evolution of novel protein functions is essential for better adaptation to different environments, explanation of this process has been a central problem of evolutionary studies.

Arrangement of protein structure is explained with several levels of organization and changes that disrupt any of these levels can have an affect on the overall protein function. The four levels of protein organization are: primary structure, which is defined by the amino acid sequence; secondary structure, defined as a regularly repeating local structure stabilized by hydrogen bonds – its most common types being alpha helix, beta sheets and turns; tertiary structure, or the overall shape of a protein, which is stabilized by non-local

interactions – hydrophobic attractions, electrostatic interactions, hydrogen and disulfide bonds, as well as by post-translational modifications; and quaternary structure, which is the structure formed by several individual protein molecules, all functioning as a part of the same protein complex. Final protein structure and function can depend on the action of other proteins in the cell, in particular when the protein depends on chaperones for folding, peptidases for activation, or specific enzymes for posttranslational modifications. However, the majority of changes in proteins are the result of mutations in the gene sequences that encode proteins. These include both – mutations that result in changes of single amino acids, but also mutations that result in larger scale changes, such as deletion, duplication or insertion of a longer stretch of amino acids.

It is important to note that many genes in higher eukaryotes do not code for one protein only. Rather, thanks to alternative splicing, they can produce several protein products. A radical example for this is neural protein Dscam that can have more than 38,000 isoforms in *Drosophila* (Wojtowicz et al., 2004). This has important implications for the studies of gene evolution, as well as studies on a single gene level, since, in order to appreciate the full repertoire of gene function, it is necessary to take into account all protein isoforms of the gene. For example, alternative inclusion of a single exon can have severe consequences for the overall function of the produced isoform.

In this introduction, I will first give an overview of the ongoing work that aims to describe functional elements in proteins and group the related elements together. I will then describe the general aspects of protein evolution and discuss the previous efforts for its systematic study. Finally, I will discuss the role of alternative splicing in creating different protein products of a same gene,

## 1.1 Characterization of functional elements in proteins

Different functional elements in proteins frequently have specific characteristics that distinguish them from other protein regions. Hence, systematic knowledge about a class of protein segments that share a similar function enables the recognition of these elements in uncharacterized protein sequences and ultimately a better understanding of protein function and regulation. In this section, I will discuss different types of protein functional elements, as well as commonly used approaches to identify these in protein sequences. Organization of functional elements in proteins defines protein architecture, and a focus of this thesis is on the changes in proteins that are the result of a gain or loss of these elements between protein homologues or different isoforms of the same gene.

### 1.1.1 Protein domains

By the standard definition, protein domains are described as basic structural, evolutionary and functional units of proteins (Holm and Sander, 1994). According to this, an individual domain is an independent folding unit in a polypeptide chain; a segment of amino acid sequence, which corresponds to a domain, is inherited and conserved in differing surrounding contexts; and distinct biological function is assigned to the domain coding segment of a protein sequence. However, dependence on structural and functional evidence restricts these well-defined domain assignments to only a handful of proteins. Therefore, a complementary domain definition, based on the sequence homology, is widely used in domain annotation.

Homology between protein regions can be identified by using pairwise sequence comparison methods, such as BLAST (Altschul et al., 1990). However, not all residues in a protein domain/family are equally well conserved. Methods that use sequence profiles were shown to be more sensitive for domain detection. These approaches rely on a multiple alignment of known members of a domain family, from which the frequency of site-specific residues are

calculated. Profile hidden Markov models (HMMs) (Eddy, 1998) formalise the more simple position specific scoring matrices (Gribskov et al., 1987), which can be used for this, into probabilistic models and allow insertions and deletion states in the models (Figure 1.1). Application of profile HMMs for domain detection has been shown to be very successful and has had a high impact on the understanding of newly sequenced genes and genomes (Bateman et al., 2002).

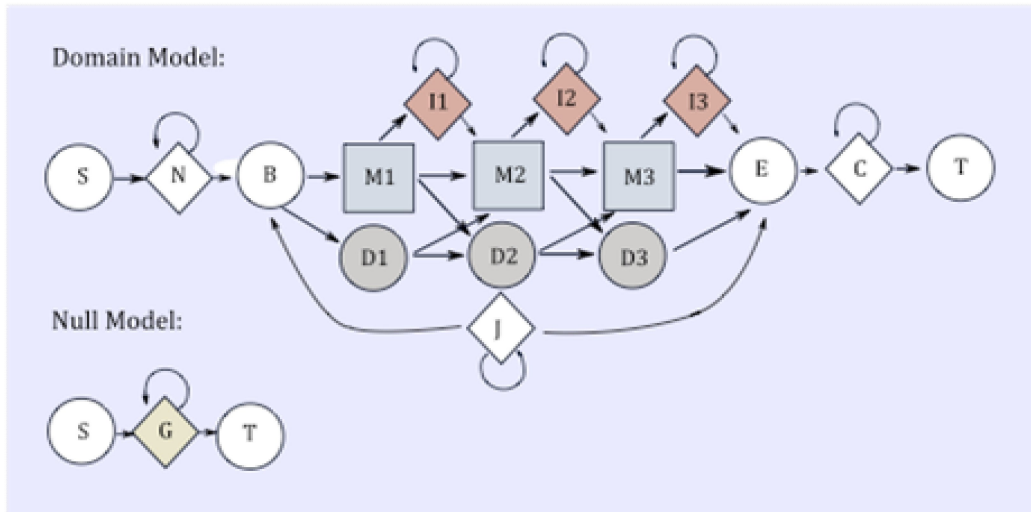


Figure 1.1: Diagram of profile hidden Markov model. States shown as squares or diamonds emit symbols, while those shown as circles do not. Each match state  $M_i$  corresponds to a column in a multiple alignment which emits over a distribution of amino acids. Insert states  $I_i$  allow for the segments of query sequence not present in the protein family and delete states allow for deletions of conserved residues in the protein family from the query sequence. The transition to the  $J$  state allows for multiple hits of the model to a single query sequence. The  $N$  and  $C$  states are analogous to insert states but occur before and after the model hit, respectively. The  $B$  and  $E$  state mark start and end of a hit to the query, while  $S$  and  $T$  are the overall start and end states. The null model emits according to a background distribution. The figure is adapted from (Coin, 2008).

The most systematically developed collection of domain models, based on profile HMMs, is the Pfam database (Finn et al., 2010), (Figure 1.2). The Pfam database is composed of two parts: Pfam-A and Pfam-B. Pfam-A is a curated section of Pfam that contains documentation and Profile-HMMs for each protein family. Manual annotation of Pfam-A families allows improvement of the initial multiple alignments and inclusion of available external information about the

proteins. Pfam-B is an automatically generated set of protein families, which is currently taken over from the ADDA database (Heger et al., 2005). Pfam-B families have no associated functional annotation and no profile-HMMs. They are in general of much lower quality than Pfam-A families, as their alignments have not been manually checked. Moreover, some Pfam-B families are composed of low complexity regions and may not reflect true relationships. Pfam domains are predicted solely from conserved sequence features. Some other databases make use of available protein structures when assigning domains to proteins. A structural classification of proteins (SCOP) database provides comprehensive description of the structural and evolutionary relationships of the proteins of known structure (Andreeva et al., 2008). The SUPERFAMILY database consists of a library of profile HMMs that represent all proteins of known structure (Wilson et al., 2009); each model in the library corresponds to a SCOP domain and aims to represent an entire superfamily. Thus, this approach enables structural assignments to protein sequences. The CATH database is also centred on domain structures, but it aims to recognize structural elements shared by different domains, as well as distantly related structures (Greene et al., 2007). The four main levels of CATH classification are protein class (C), architecture (A), topology (T) and homologous superfamily (H). Class describes the secondary structure composition of each domain, architecture the shape revealed by the orientations of the secondary structure units, such as barrels and sandwiches. At the topology level, sequential connectivity is considered, such that members of the same architecture might have quite different topologies. When structures belonging to the same T-level have suitably high similarities combined with similar functions, the proteins are assumed to be evolutionarily related and put into the same homologous superfamily. Gene3D assigns structural domains from the CATH database to whole genes and genomes (Yeats et al., 2008). Matches to structural domains are found using the PSI-Blast (Altschul and Koonin, 1998). Two automatically generated databases that cluster protein domains are the ProDom (Bru et al., 2005) and ADDA databases. ProDom iteratively invokes PSI-Blast to cluster protein domains, and ADDA Automatic Domain Decomposition Algorithm. This algorithm first aligns representative protein sequences with BLAST (Altschul et al., 1990), splits them into domains and then organizes these

domains into protein domain families. Other domain databases that use HMMs for domain classification are SMART (Letunic et al., 2006) and TIGRFRAM (Haft et al., 2003). The SMART (Simple Modular Architecture Research Tool) database is focused on certain types of domains, such as extracellular and signalling domains, while TIGRFRAM strives for broad coverage of microbial proteins. The Prosite database consists of a library of profiles and patterns that describe protein domains, families and functional sites (Hulo et al., 2006). The PRINTS database is a collection of nonoverlapping motifs for the identification of family members (Attwood et al., 2003). The motifs are derived from ungapped multiple sequence alignments that help to identify the most conserved regions of the protein family. Prints families tend to be more specific and are useful for detecting subfamilies. The BLOCKS database contains blocks, i.e. ungapped multiple sequence alignments, for each family (Henikoff et al., 2000). These are equivalent to the motifs in the PRINTS database, and in fact the families in BLOCKS are currently derived from Prosite and Prints families. Finally, InterPro is an integrated database - a result of collaboration between different domain family databases and the UniProt Knowledgebase (Hunter et al., 2009). The goal of this collaborative project is to have a centralized resource for protein classification and automatic annotation.

Presence of an already described domain in protein sequence is one of the most informative indications of protein function. Therefore, protein domains are used as the basis for automatic protein functional classification and annotation. Presence of other functional elements in a protein sequence can also aid in better understanding of protein's role in a cell. In the following text, I discuss the function of, and methods to characterize, disordered regions and posttranslationally modified sites in proteins. When disordered regions are conserved, it is possible that they are also classified as protein families, so protein domain annotations can overlap with disordered segments in proteins. However, these segments are crucially distinct from standard protein domains - both from the aspect of structure and function. Other classes of functional elements in proteins, such as transmembrane regions, or signal peptides, are also well described and methods for their detection are in use (Kall et al., 2004), but I don't address them here separately.



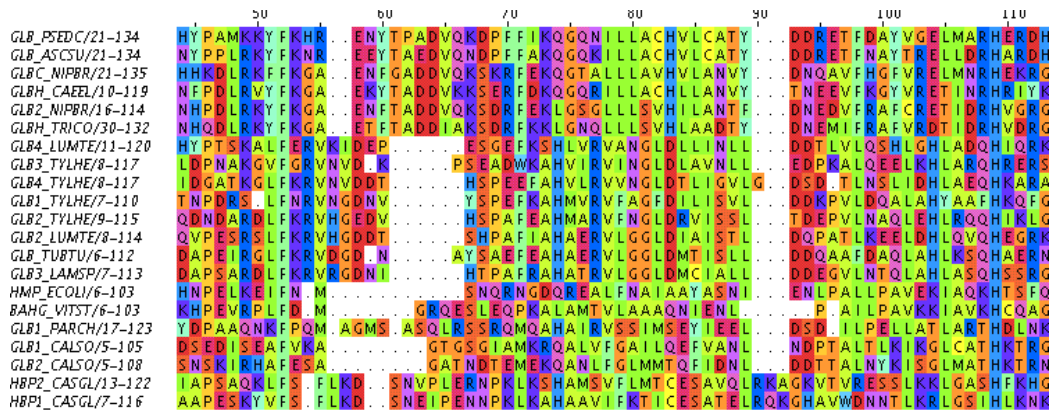


Figure 1.2: An example of a seed multiple alignment for the Pfam Globin family (Pfam accession: PF00042). The seed alignment is used to build an HMM model of a family, which is the used for identifying the same domain in other proteins.

### 1.1.2 Disordered protein regions

Intrinsically unstructured, or disordered, regions in proteins are characterized with the lack of stable secondary and/or tertiary structure (Dunker et al., 2001; Dyson and Wright, 2005). In some cases, though, disordered segments can adopt a fixed three-dimensional structure after binding to other macromolecules in a cell, as exemplified with DNA binding domains of different transcription factors (Gspomer and Babu, 2009). The discovery of proteins that are unstructured over their whole length challenged the traditional view that a well-defined structure is required for correct protein function. Moreover, further work demonstrated that the flexibility of disordered residues actually provides these proteins with specific functional benefits. The functional importance of protein disorder is underlined with the observations that disordered proteins commonly play a role in signal transduction, cell-cycle regulation, gene expression and chaperone activity (Tompa, 2005; Wright and Dyson, 1999).

Experimentally, the lack of a stable tertiary structure in proteins is usually demonstrated by using solution-state NMR, circular dichroism, fluorescence spectroscopy and small angle X-ray scattering measurements (Gspomer and Babu, 2009). The database DisProt (Vucetic et al., 2005) is a repository of proteins with experimental evidence of a lack of structure. In addition to this, since disordered protein segments have a distinct amino-acid

composition, they can also be predicted from protein sequence. Disordered regions tend to be enriched in hydrophilic and charged amino acids that do not tend to form stabilizing interactions with other neighbouring amino acids; Alanine, Arginine, Glycine, Glutamine, Serine, Proline, Glutamic acid and Lysine (Tompa, 2005). Specific properties of disordered segments have been differently applied in disorder prediction methods. These methods can generally be classified into those that apply machine-learning approaches and use known disordered proteins for training, and those that predict disorder just from sequence properties. PONDR (Garner et al., 1998), Disopred (Ward et al., 2004), and DisEMBL (Linding et al., 2003) are examples for the former class of methods and IUPred (Dosztanyi et al., 2005) and SEG (Wootton, 1994) for the latter – SEG actually predicts low complexity regions which can serve as a good indication of disorder.

The functional classification of disordered protein regions, as explained here and as shown in Figure 1.3, is adapted from the classification suggested by Peter Tompa (Tompa, 2005). Disordered proteins or protein segments can be divided depending on whether their function results from the entropic properties of disordered chains or from the ability to flexibly bind other partner molecules. Examples for the former one are Phe-Gly (FG) disordered repeat regions of nucleoporins that regulate transport through nuclear pore complex via spatial exclusion (Denning et al., 2003), or the microtubule-associated protein 2 (MAP2) repeat domain that provides spacing in cytoskeleton (Ludin et al., 1996). Disordered regions or proteins that interact with other molecules can be further divided in those that achieve the interactions through permanent binding and those that bind their partners only transiently. Those that bind the partner molecules permanently are usually inhibitors of different enzymes, take part in different cellular complexes as assemblers, or, if partner molecules are small ligands, regulate the ligand dynamics. Disordered regions and proteins, which form only transient interactions, do that either by exposing flexible binding sites, such as those for posttranslational modifications, or they function as protein or RNA chaperones (Tompa and Csermely, 2004).

Comparison between fractions of disorder in proteins from fully sequenced representative genomes from the three kingdoms of life revealed a

significant increase of native disorder between eukaryotic genomes compared to archean or eubacterial genomes (Ward et al., 2004). Moreover, among eukaryotes the fraction of disorder increases with organism complexity (Haynes et al., 2006). In eukaryotes, disorder is especially abundant in hub proteins, i. e. in proteins with a high number of interaction partners (Dosztanyi et al., 2006; Haynes et al., 2006). In line with this, independent studies reported that cancer-associated and signalling proteins are also enriched in disorder (Iakoucheva et al., 2002). Furthermore, there are indications that contacts between two disordered regions might be the most frequent type of interactions in the protein-protein interaction network (Shimizu and Toh, 2009). Hence, disordered proteins are suggested as attractive novel drug targets (Cheng et al., 2006).

The benefit of using disordered regions in protein interactions is most obvious when binding sites are exposed for transient interactions, such as sites of post-translational modifications. Disordered segments can be easily accessed by modifying enzymes which add or remove a modification, and by effector proteins which are regulated by the (un)modified proteins (Gsponer and Babu, 2009). Easy accessibility of these sites enables precise time regulation of a process. Therefore, it is not surprising that disordered regions in proteins frequently contain short linear peptide motifs (Neduva and Russell, 2005) that are important for protein function and recognized by specific protein partners. The most comprehensive collection of described linear motifs - small functional sites in proteins - is catalogued in the Eukaryotic Linear Motif (ELM) database.

Disordered proteins are more sensitive to proteolytic degradation and have a short lifetime (Tompa, 2005; Wright and Dyson, 1999). Moreover, the abundance of disordered proteins is additionally controlled on the level of regulation of transcript clearance and translational rate (Gsponer et al., 2008). Thus, both life-span and synthesis of these proteins seem to be finely regulated. Rapid turnover is a desirable characteristic of proteins involved in cell cycle regulation and in transcriptional and translational processes. These exactly are the functional categories that disordered proteins are enriched in (Tompa, 2005; Wright and Dyson, 1999). Therefore, the intrinsic characteristics of disordered proteins make them especially adapted to the roles they perform in a cell. This ensures that they are available in appropriate amounts and only during a short

time interval (Gspomer et al., 2008). Moreover, disordered proteins that form transient interactions and are readily accessible for protein modifications provide another advantage for usage in finely regulated signalling pathways.

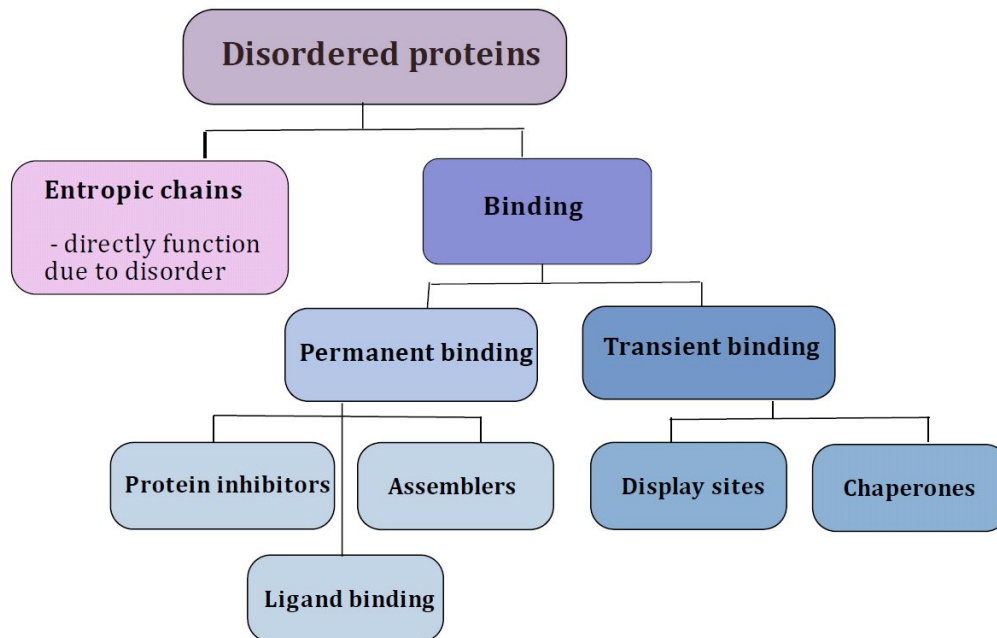


Figure 1.3: Functional classification of disordered proteins. Examples of disordered proteins from each category are described in the text. Illustration is adapted from Tompa (2005).

### 1.1.3 Sites of posttranslational modification

Posttranslational modifications (PTMs) are covalent processing events that modify proteins. These modifications rely on the activity of other proteins – enzymes, which either proteolytically cleave the protein or add a modifying group to its amino acid(s). The majority of eukaryotic proteins undergo posttranslational modifications, which modulate their activity (Mann and Jensen, 2003). PTMs can modify stability, activity state, localization or turnover of a protein, as well as its interactions with other proteins (Mann and Jensen, 2003; Walsh, 2006). Even though protein modification is a widespread phenomenon which regulates numerous aspects of protein function, only a small subset of all PTM sites has been discovered (Olsen et al., 2006). This is exemplified with protein phosphorylation, which is the most intensively studied type of protein PTM, and estimated to affect about one-third of all proteins (Cohen, 2001). However, currently only a small fraction of protein PTM sites are described (Olsen et al., 2006). Development of mass spectrometry methods, which provide enough sensitivity for large-scale studies, offers great promise in scaling up detection and our understanding of different PTMs (Mann and Jensen, 2003).

Protein PTMs are used in numerous cellular processes. Proteolytic cleavage is important for the activation of many proteins; these are firstly synthesised as inactive precursors that are later on activated through limited proteolysis. Examples for this are pancreatic enzymes and enzymes involved in blood clotting (Neurath and Walsh, 1976). Phosphorylation is particularly important in signalling, where kinase cascades are regulated by reversible addition and removal of phosphate groups (Mann and Jensen, 2003). Similarly, ubiquitination plays an essential role in the cell cycle where it marks cyclins for destruction at defined time points (Mann and Jensen, 2003). Methylation and acetylation can both modify the activity of histones and hence regulate gene expression (Rice and Allis, 2001). Addition of fatty acids, such as palmitoyl or myristoyl, is used to promote membrane binding and target proteins to specific organelles (Resh, 1999). Glycosylation is used both in signalling (Haines and Irvine, 2003) and in defining proteins that are excreted or exposed on a cellular surface (Gahmberg and Tolvanen, 1996).

PTM sites frequently reside in disordered protein segments (Fuxreiter et al., 2007). Advantages of this are discussed above in the text. In particular, protein phosphorylation has been strongly linked to intrinsically disordered protein segments (Iakoucheva et al., 2004). Since these regions evolve rapidly, and phosphosites are relatively short, it has been suggested that some of the annotated sites are not functional, and that the process of signal transduction tolerates a certain level of noise (Landry et al., 2009). Moreover, phosphosites of known function are significantly more conserved than those of unknown function, and hence it has been suggested that evolutionary conservation could give an indication of the actual functionality of a phosphosite (Landry et al., 2009). However, studies on yeast have suggested that the position of most phosphorylation sites is not conserved in evolution and that clusters of sites tend to shift positions in rapidly evolving disordered regions, which could also be the mechanism for the faster evolution of kinase-signalling circuits (Holt et al., 2009).

## 1.2 Protein evolution

Evolutionary footprints are evident in protein sequences, where in general the level of sequence divergence reflects divergence times between organisms. Hence, present day protein sequences, together with ribosomal sequences, are often used to assign organisms to their phylogenetic groups (Feng et al., 1997). Additionally, divergence in protein sequences represents a molecular clock, which, after calibration with the available fossil record, can be applied to estimate divergence times between more distant organisms (Feng et al., 1997). However, it is important to note that protein sequence divergence is not a random evolutionary process, but mutation patterns are largely shaped by proteins structural and functional constraints. Even a single point mutation in a protein can have a dramatic effect on the protein function. For example, amino acids in an enzyme's active site are usually highly conserved and their mutations can completely abolish the original function. Sometimes, substitutions of the active-site residues can lead to catalytically inactive forms that can later adopt

new functions, such as those in regulatory processes (Pils and Schultz, 2004). Additionally, mutation in an enzyme's catalytic site can adapt its specificity to a different substrate, and there are examples of enzymes that have evolved to catalyse different reactions on the same structural scaffold using this mechanism (Bartlett et al., 2003).

When a protein is folded into a stable structure, mutations in the primary sequence introduce a risk to its structural stability. The first level of protein structural hierarchy is defined with elements of secondary structure, and the next higher level – protein fold – with the arrangement of secondary structure elements. Examples of protein folds are helix bundle, which is a fold composed of several alpha helices; beta-barrel, which is a large beta-sheet that forms a closed structure; and Rossman fold, which is a fold composed of interchanging beta strands and alpha-helices, commonly found in nucleotide-binding proteins. Interestingly, analysis of known structures suggests that the total number of folds in nature is limited (Chothia, 1992; Goldstein, 2008). Moreover, some folds are extremely common while other folds are shared only between a few related proteins (Goldstein, 2008). A possible explanation for this is that folds that are suitable for common functions in cells, or for a wider range of different functions, have been most often adopted in evolution (Goldstein, 2008). As a consequence of this, the introduced mutations are likely to disrupt the structural stability. Additionally, many other factors - apart from protein structure and function - affect protein evolution. Other genomic factors that play an important role are: positions of the encoding genes in genomes, gene expression patterns, protein positions in biological networks (Pal et al., 2006) and also availability of buffering mechanisms, such as chaperones, which can stabilize intermediate, slightly deleterious, protein mutations (Tokuriki and Tawfik, 2009). Apart from experiencing mutations on the amino acid level, whole genes encoding proteins can be gained or lost during evolution. Gains can occur either through exonisation of non-coding sequences, or through gene duplications – discussed below. Gene propensities to be lost, similarly to the mutation propensities of protein amino acid sequences, depend on their essentiality for the organism, level of expression and a number of interaction partners (Krylov et al., 2003). Finally, another principal mechanism of protein evolution is domain shuffling.

The unit of evolution here is a protein domain and, hence, the changes in proteins are of larger scale than those observed in amino acid divergence. In the following section, I will discuss reports from the studies on how new domain combinations are formed, and what role they play in protein and organism evolution.

### 1.2.1 Domain shuffling

Above in the text, I introduced the terms 'protein fold' and 'protein domain'. When sequences with the same fold are evolutionary related, and the protein domain is structurally independent from the rest of the protein, fold and domain definitions overlap. In my thesis, I focus on protein domains and their roles as independent evolutionary units. The majority of proteins consist of at least two domains, and many domains can occur in combinations with different domain partners. Thus, multidomain proteins are frequently created through rearrangements between domains (Moore et al., 2008). Since the same domains are reused in different combinations, domain duplication is an important prerequisite for novel domain rearrangements. The majority, i.e. 98%, of domains in humans are present in at least two copies in the genome (Chothia et al., 2003). Additionally, when the same domain combination, i.e. two or more domains, are present in two otherwise non-homologous proteins, domain order is conserved in more than 90% of the instances (Vogel et al., 2004). This implies that these regions share a common ancestor and underscores the role of domain duplication in creation of novel multidomain proteins.

Observed domain combinations are only a small fraction of all possible combinations (Chothia et al., 2003). This shares a similarity with the evolution of protein folds and suggests that protein evolution could be affected by functional and structural constraints on all levels. In line with this, analysis of experimentally characterized protein structures of multidomain proteins reported that independent folding of structured domains can be achieved through loosely packed or small interfaces between the domains (Han et al., 2007). Another observation from the studies of multidomain proteins is that domains that occur most often in the genomes also have many different



combination partners (Vogel et al., 2005). Interestingly, these domains are often shared between members of larger phylogenetic groups. Study of domains with known structure (Chothia et al., 2003) showed that domains that are shared between all eukaryotes or all animals make more than 80% or 95%, respectively, of domains in the human genome. A significant fraction of this is a result of lineage-specific expansions of some of the shared domains (Chothia and Gough, 2009).

Similar domain architectures are usually explained with shared ancestry and convergent evolution is considered to be rare (Apic et al., 2001; Gough, 2005). Studies of rearrangements in the evolution of multidomain proteins have shown that the evolution of the majority of multidomain proteins can be explained with insertions and deletions of domains from protein termini (Bjorklund et al., 2005; Weiner et al., 2006), with the exception of domain repeats, where the changes in the number of domains also occur in the middle of proteins (Bjorklund et al., 2006). These studies were performed by comparing proteins with similar, but not identical, domain assignments. However, domain architectures can also be used to build evolutionary trees, which can be useful when frequent domain rearrangements make it difficult to recognize related proteins from the amino acid level. This method has been used in a number of studies for inferring phylogeny - covered in the review by Moore and colleagues (Moore et al., 2008), and tools for finding related proteins based on domain architecture are also available (Geer et al., 2002; Storm and Sonnhammer, 2001). A recent study used a tree based on the distances between domain architectures from all species with good quality genomes as a guide in the study of evolution of multidomain proteins (Ekman et al., 2007). Mapping the changes in multidomain proteins to species divergence times showed that the major changes in domain architectures have occurred in the process of multicellularization and then within the metazoan lineage (Ekman et al., 2007). This suggests that accelerated formation of novel domain architectures was needed for the emergence of novel, more complex traits. Jin and colleagues propose that changing combination partners relieves the pressure for a domain to maintain the original function and allows it to acquire an entirely new intrinsic function (Jin et al., 2009), as illustrated in Figure 1.4. This can expand the function of an original protein and

modify the cellular process that this protein is involved in. Frequently, domains with a number of different domain partners are involved in signalling and it was suggested that shuffling of these domains was a crucial step in the evolution of complex cellular networks (Pawson, 2003). Similar to this, the distinguishing feature of the proteomes of multicellular eukaryotes is a high fraction of domain repeats (Ekman et al., 2005). Domain repeats often have a role in protein-protein interactions or binding to other ligands (Bjorklund et al., 2006). Thus, this could be another category of domain architecture rearrangement events that was important for the development of complex intra- and intercellular networks and subsequently for the evolution of novel phenotypic traits in the metazoan lineage.

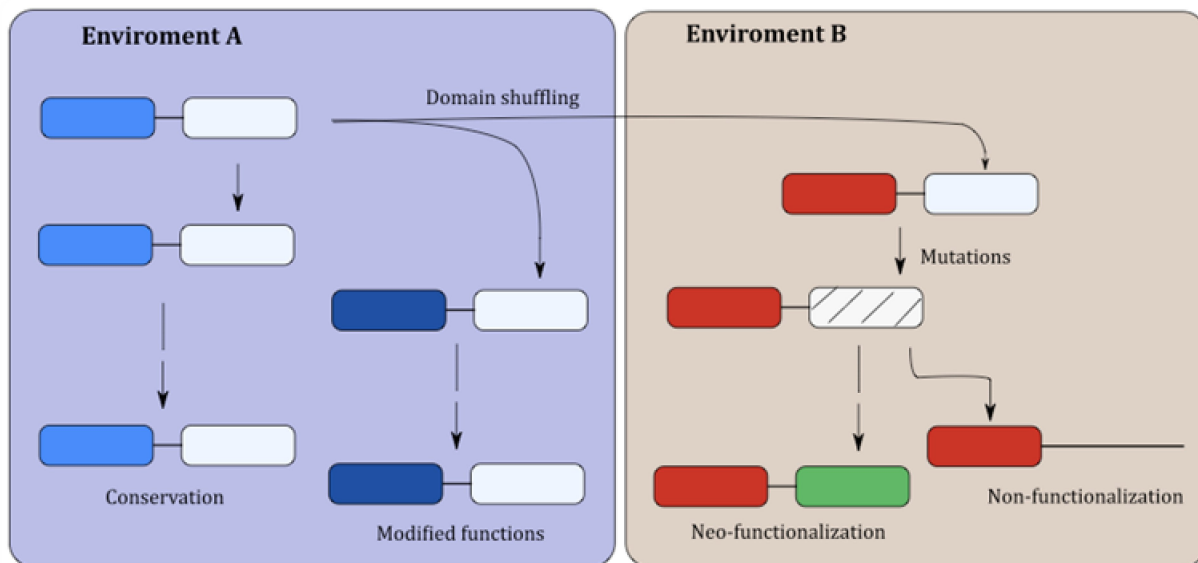


Figure 1.4: Domain shuffling and domain evolution. When domain shuffling changes the environment of a domain, the domain is likely to experience more radical changes in sequence and function. The domain environment is defined by the subcellular localization and interaction partners of a domain. The figure is adapted from Jin et al. (2009). If, through shuffling, a domain is attached to a protein that has similar interaction partners and localization as the ancestral protein that the domain was a part of (left panel in the figure), domain sequence and function evolve more slowly than if the domain is attached to a protein that operates in a different cellular compartment and/or has different protein partners (right panel in the figure) compared to the ancestral protein.

Several studies focused on specific examples of domain shuffling and demonstrated its importance in the development of complex systems or evolution of signalling pathways. One of these studies investigated the role of domain shuffling in the evolution of vertebrates (Kawashima et al., 2009). The evolution of vertebrates included a number of important and novel events, such as the development of cartilage, the immune system and craniofacial structures (Kawashima et al., 2009). The study showed that proteins which are components of vertebrate-specific structures, such as cartilage and the inner ear, had novel domain combinations, thus suggesting that domain shuffling made a strong contribution to the evolution of vertebrate-specific traits (Kawashima et al., 2009). An interesting example from the study is the Xlink domain in the aggrecan protein, which is one of the major components of cartilage. This domain appears to be recruited in the cartilage matrix protein by domain shuffling, while in protochordate ancestors, Xlink was most likely used as a surface molecule of blood cells (Kawashima et al., 2009). An example of a cellular pathway where domain shuffling played an important role is the Notch signalling pathway. This pathway regulates cellular identity, proliferation, differentiation and apoptosis, and plays an important role in development (Gazave et al., 2009). Systematic study of genes involved in this pathway in a number of eukaryotic species showed that this pathway is specific to Metazoans, and moreover, that the origin of several components of the pathway occurred through shuffling of pre-existing domains (Gazave et al., 2009).

Research that puts domain shuffling in context with other types of protein evolution – point mutation and protein duplication - suggests that this is the most powerful source for innovation of gene function (Conant and Wagner, 2005). Experimental evolutionary studies show that function evolves at a much faster rate following domain rearrangements than following point mutations (Leong et al., 2003; Powell et al., 2000) or gene duplications (Peisajovich et al., 2010). The incidence of domain shuffling in eukaryotes is reported to be significantly less frequent than gene duplication events (Conant and Wagner, 2005). However, evolution by domain shuffling is most likely closely linked to other types of protein evolution: there is evidence that domain shuffling relies on gene duplication, which provides domain copies for shuffling (Vogel et al., 2005),

and after new domain combinations are formed, point mutations in the shuffled domain can occur with a higher frequency than in the original domain context (Jin et al., 2009).

### 1.2.2 Mechanisms for formation of novel genes

Domain shuffling is a powerful mechanism for protein evolution. However, a change in a protein that we observe as domain shuffling could be a result of different gene rearrangement mechanisms. Comparisons of protein domain architectures can only give indications on which mechanisms could have caused the observed changes (Bjorklund et al., 2005; Weiner et al., 2006). On the other hand, studies on the origins of new genes are primarily focused on mechanisms that underlined the emergence of novel genes and functions (Long, 2001). The two approaches to a study of evolution of novel functions are complementary to each other; mechanisms that underlie the evolution of novel genes could have also caused changes in protein domain architecture, and alternatively – gain or loss of a protein domain is a strong indicator of a change of function during gene evolution. Here, I cover recent work that addressed emergence of novel protein coding genes and discuss which of the underlying mechanisms could have also played a role in domain shuffling.

The main interest in studying the occurrence of novel genes, and underlying mechanisms for it, comes from a notion that novel genes might have played a significant role in the evolution of lineage- or species-specific traits (Kawashima et al., 2009; Khalturin et al., 2008). A powerful mechanism that can lead to the evolution of novel functions is gene duplication. The role of gene duplications in evolution of novel traits has been debated for more than four decades (Ohno, 1970) and I discuss it as a separate aspect of gene and protein evolution in the next section. Next, recombination of either duplicated or single copy genes can result in the creation of proteins with novel domain arrangements. The two best-studied means of recombination are non-allelic homologous recombination (NAHR, Figure 1.5) (Hurles, 2004) and non-homologous end joining (NHEJ) (Arguello et al., 2006). These mechanisms recruit different proteins (Haber, 2000) and differ in whether they require short

regions of sequence similarity for their action or not; NAHR, unlike NHEJ, acts between the short blocks of high identity sequences. These blocks could have originated through previous duplications of genetic material, or even through expansion of transposons in the genome (Babushok et al., 2007). An example of a gene that evolved through DNA recombination is the Hun gene in the *Drosophila* lineage (Arguello et al., 2006). This gene is a partial duplicate of Baellchen gene, from another chromosome, and after its duplication it has recruited intergenic sequence and evolved independently in each *Drosophila* species. A lack of obvious direct repeats around the duplicated region led the authors to propose that the underlying recombination mechanism was NHEJ (Arguello et al., 2006). Another example is a primate-specific chimeric gene family that expanded as a result of intrachromosomal segmental duplications, and was derived through joining of exons from the RanPB2 gene with exons from the neighbouring GCC2 gene, which code for the GRIP domain (Cicarelli et al., 2005). RanBP2 is the largest protein found in the nuclear pore complex, while the GRIP domain has been shown to be sufficient for targeting to Golgi. The new chimeric protein - named RGP (for RanBP2-like, GRIP domain containing protein) - was indeed found to localize inside cytoplasmic regions, while the ancestral RanPB2 protein is almost exclusively found at the nuclear envelope (Cicarelli et al., 2005). Emergence of this chimeric protein is closely connected to segmental duplications of the RanBP2 gene in primates. The observed intrachromosomal duplications could have occurred through NAHR, which more frequently acts between the regions on the same chromosome (Arguello et al., 2006). However, the birth of the RGP gene also required joining of exons from two adjacent genes, and this supports the theories that intergenic splicing could play an important role in assisting gene fusions in eukaryotes (Babushok et al., 2007).

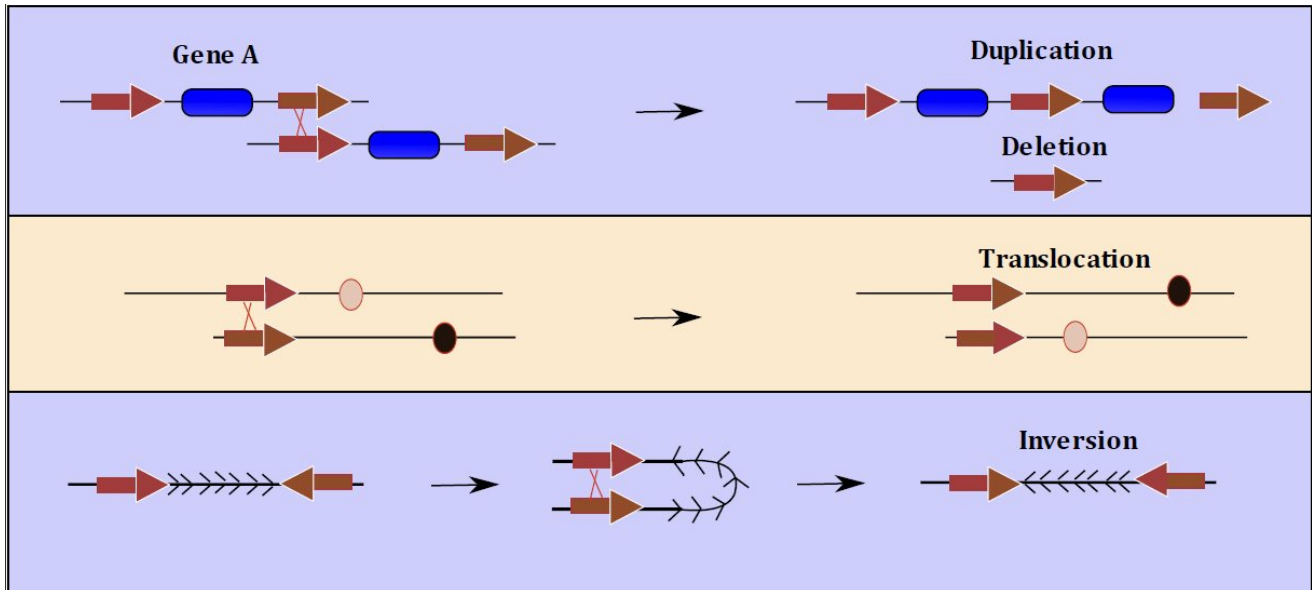


Figure 1.5: Possible effects of Non-allelic homologous recombination (NAHR) on genome evolution. NAHR between two highly similar segments in the genome can cause different types of rearrangements, depending on the location and orientation of these segments. Thus, NAHR between adjacent duplicated sequences can result in tandem duplications and deletions (top figure). When the similar segments are on different chromosomes NAHR can result in translocation (middle figure), and intrachromosomal recombination between inverted similar segments can result in inversions (bottom figure).

In prokaryotes, the dominant mechanism for domain gains is fusion of adjacent genes (Pasek et al., 2006). However, more complex gene structures in eukaryotes make simple fusion of coding sequences less likely. So far, there is one example for this in the literature (Ponce and Hartl, 2006). Sdic is a new gene in *Drosophila melanogaster* that arose after its ancestral genes Cdic and AnnX, that are next to each other in the genome, were duplicated. This was followed with several deletions that eliminated regions between the two gene copies in the middle – in the order AnnX and Cdic - and fused them into a chimeric Sdic gene, as illustrated in Figure 1.6. Even though such scenarios are likely to be rare in the evolution of eukaryotic genes, there are other mechanisms which can assist fusion of adjacent genes with complex structure. Intergenic splicing was observed to be relatively frequent in mammalian genomes. By this mechanism, novel chimeric proteins can be created. It was suggested that when new proteins are advantageous for the organisms they are created in, mutations inside the regulatory regions that distinguish expression of two different genes will be selected for and the chimeric product will be also fixed on the gene level (Babushok et al., 2007). An example for this is a fusion of two adjacent human genes, KUA and UEV (Thomson et al., 2000). The resultant intergenic transcript skips the exons with stop and start codon between the two originally separate genes to ensure successful translation of a final product. Interestingly, KUA and UEV were most likely also initially juxtaposed as a result of a recombination event.

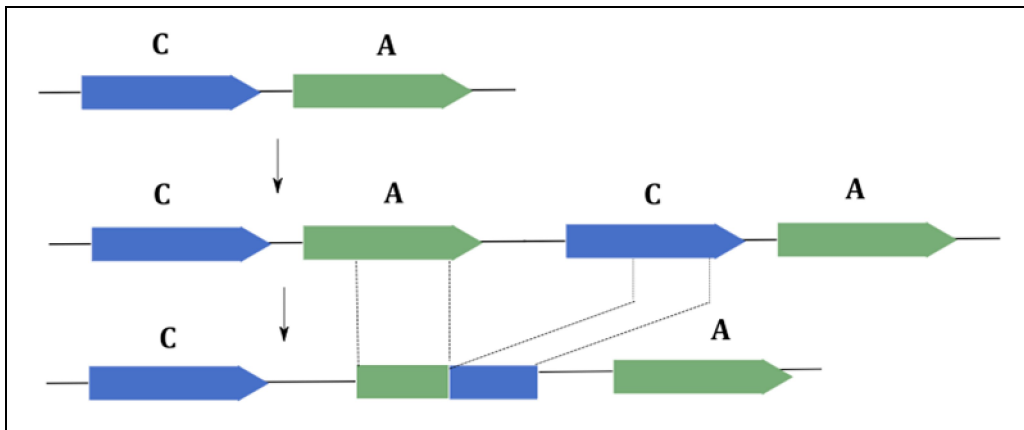


Figure 1.6: Example of a chimeric gene formed by gene fusion. The model is a simplified scenario of the evolution of the Sdic gene. Steps in the evolution of this gene include tandem duplication of neighbouring genes named C and A. This is followed with the deletion of parts of genes A and C as well as intergenic regions between them which results in the fusion of two partial coding regions. Finally, later evolutionary events include the emergence of new start and stop codons and recruitment of regulatory elements of the new gene.

Another mechanism that can underlie evolution of novel proteins is retroposition. Retrotransposons, such as for example LINE1, expand in the genome by reversely transcribing their own mRNA and inserting a copy randomly in the genome (Babushok et al., 2007). However, their machinery can also be used to reversely transcribe cellular mRNA, and that is the mechanism for the emergence of processed pseudogenes. Additionally, only portions of cellular mRNA can be transcribed, or templates can be switched during transcription, thus resulting in combination of different cellular mRNAs, or cellular mRNA and a transposable element (Babushok et al., 2007). Furthermore, this mechanism can fix mRNAs created by intergenic splicing as novel genes. One such example is the emergence of the gene PIPSL in primates, which combines the lipid kinase domain of PIP5K1A and the ubiquitin-binding motifs of PSMD4 – its two ancestral genes (Babushok et al., 2007a). PIPSL is reported to have experienced strong positive selection, and is found to be transcribed specifically in the testes (Babushok et al., 2007a). Testis is in general a more permissive environment for gene expression, and the organ where young retrogenes can be found expressed (Betran et al., 2002). Because of that, testis has been proposed



as a tissue where accelerated evolution of genes takes place, assuming at the same time that the newly evolved genes can later adapt to other tissues (Kaessmann et al., 2009).

Retrotransposons, together with retroviruses and other parasitic elements in the genome, can contribute to gene evolution also by directly incorporating into the other genes in the genome (Deininger et al., 2003). It has been reported that new exons can arise through exonisation of Alu elements or other parasitic elements in the genome (Sorek and Ast, 2003; Sorek et al., 2004). An important example identified in these studies is the ADAR2 enzyme – a double-stranded RNA-specific adenosine deaminase that is involved in the editing of mammalian messenger RNAs by site-specific conversion of adenosine to inosine (Rueter et al., 1999). This enzyme contains 40 amino acids in its active site that are derived from an Alu element. This addition changes the activity of the enzyme essential in mammals. Another example is the incorporation of a DNA transposon into a cellular gene which gave rise to the ZBED6 transcription factor in eutherians (Markljung et al., 2009). ZBED6 has an important role in the regulation of muscle growth, and might affect the expression of numerous genes involved in other biological processes (Markljung et al., 2009). An example of genes that evolved from retroviruses are syncytin genes, which stem from the envelope genes of endogenous retroviruses and have evolved in mammals (Mi et al., 2000). Importantly, syncytin genes play key roles in placentation.

Evolution of novel protein coding genes was long believed to be strongly linked to gene duplication (Ohno, 1970) and the probability that new functional proteins are created de novo was argued to be extremely unlikely (Jacob, 1977). In line with this, it was noted that novel folds that are created during evolution can be presented as modified topological combinations of already known motifs of secondary sequence (Fernandez-Fuentes et al., 2010). Hence, recent reports of protein coding genes that have evolved completely from scratch were rather surprising. One example for this is morpheus gene family, that evolved in primates, and after its birth has experienced a series of segmental duplications and positive selection in hominoids (Johnson et al., 2001). Studies in *Drosophila* also reported 14 de novo-originated genes (Levine et al., 2006; Zhou et al., 2008). Finally, three de novo human specific genes were recently reported (Knowles and

McLysaght, 2009). Comparison of these genes with related, non-coding sequences in other primates revealed mutations that allowed formation of functional open reading frames, and available protein evidence proved that these genes are indeed translated. Interestingly, two out of the three human-specific genes fall within introns of the genes on the opposite strand. This suggests that possibly transcription of the genes on the opposite strand and open chromatin structure permits transcription of the de-novo genes even without the presence of sophisticated regulatory signals (Siepel, 2009). Therefore, if whole genes can evolve from previously non-coding regions, this also implies that novel domains - fractions of coding genes - could also originate from scratch during evolution. Nonetheless, this is more likely to be the mechanism for emergence of domains defined on the basis of sequence conservation rather than emergence of novel structural units. Alternatively, novel domains can be created through point mutations of already existing domains, and hence, lineage-specific domains that hence contribute to novel domain arrangements, are likely to be of both sorts.

Finally, exon shuffling has often been referred to as a separate mechanism of gene evolution (Long, 2001; Long et al., 2003). However this phenomenon is in fact a result of an already described mechanism - recombination events and possibly retroposition. Exon shuffling is a term that could include any novel combination of exons, but was frequently associated with insertions of novel middle exons that encode protein domains (Patthy, 1996), and hence is now also often used in that context (Marsh and Teichmann, 2010).

### 1.2.3 Gene duplication and protein evolution

As already stated in the previous section, gene duplication is believed to be the strongest driving force behind the evolution of novel functions (Ohno, 1970). The rationale behind this is simple; the majority of mutations are deleterious, and since, in general, each gene has evolved a specific role in the organism, disruption of gene function in parallel affects the organism fitness. However, when a gene is duplicated, it is theoretically possible that one copy evolves freely and goes through intermediate stages that change its original function - as long as this does not interfere with the function of the other copy. Gene duplicates can

be created through recombination or retrotransposition events, or as a result of chromosome or whole-genome duplications (Zhang et al., 2003). Similarly, duplicate genes in the human genome originated mostly from one or two rounds of whole genome duplication before the divergence of vertebrates, subsequent smaller segmental duplications (Gu et al., 2002) and more recent expansion of retrogenes (Kaessmann et al., 2009). Interestingly, gene survival is dependent upon the mechanism of duplication. For example, duplication of a single gene that is a part of protein complexes or is involved in signalling processes can disrupt the dosage balance in the cell. Therefore, duplicates of such genes are underrepresented in the genomes (Makino and McLysaght, 2010). On the contrary, after whole genome duplications, dosage-sensitive genes are present in two copies. Hence, losing a dosage-sensitive gene disrupts the newly created dosage balance and is likely to be selected against.

Genes duplicated through retroposition lack regulatory elements – since only their mRNA has been duplicated (Kaessmann, 2009). However, a surprisingly large number of such retrogenes are found to be transcribed (Zheng et al., 2005). One means of transcription could be usage of the open chromatin state and regulators of nearby genes (Kaessmann et al., 2009). Moreover, specific examples have been described where a gene after retroposition evolved a novel, positively selected, function. An example is the duplication of the enzyme glutamate dehydrogenase (GDH) (Burki and Kaessmann, 2004). GDH is important for the recycling of glutamate during neurotransmission. In humans, this enzyme exists as a ubiquitously expressed form GLUD1 and as a brain-specific form GLUD2. Interestingly, GLUD2 originated by retroposition of GLUD1 in the hominoid ancestor and went through a period of positive selection during which it acquired changes necessary for its brain-specific function. Another example for the possible effect of gene retroposition is the impact of a retrocopy derived from a growth factor gene (*fgf4*) in several common dog breeds, where this extra gene copy is solely responsible for a short-legged phenotype (Parker et al., 2009). The resulting phenotype seems to be consequence of gene dosage alteration.

Many fixed duplicated genes acquire mutations that make them non-functional over time; they become pseudogenes, and are often deleted from the

genome (Zhang, 2003). It has been proposed that important processes that lead to retention of duplicate genes in the genome are neofunctionalization and subfunctionalization (Roth et al., 2007). Neofunctionalization, or the origin of new function, is a particularly important aspect of gene evolution after duplication. Proteins with new functions underline the emergence of novel phenotypic traits, and adaptation of the function of an already existing protein to a new context is a much faster means of evolution than creation of a protein *de novo*. An example for the adaptation of gene function after duplication is the creation of the red- and green-sensitive opsin genes in humans and Old World monkeys (Yokoyama and Yokoyama, 1989). After gene duplication in this primate lineage, the two opsin proteins have diverged in function, which resulted in a 30-nm difference in the maximum absorption wavelength and enabled a sensitivity to a wider range of colours. In addition, a duplicated gene can also evolve an entirely new function. One example for this is another gene duplication event in the ancestors of humans and Old World monkeys. This duplication resulted in another gene in the RNase A gene family – eosinophil cationic protein (ECP), which after duplication went through accelerated evolution (Zhang et al., 1998). As a result, the encoded protein experienced multiple changes of its amino acids compared to the progenitor eosinophil-derived neurotoxin (EDN) protein and developed novel antibacterial activity, which seems to be independent of the ribonuclease activity (Rosenberg, 1995). During subfunctionalization, each daughter gene adopts part of the function of the parental gene (Force et al., 1999). One form of subfunctionalization is the division of gene expression after duplication (Force et al., 1999). An example for this is a pair of transcription factors, engrailed-1 and engrailed-1b in zebrafish, which are expressed in different tissues, while their mouse orthologue is present in a single copy and is expressed in all the tissues where either engrailed-1 or engrailed-1b is found in zebrafish (Force et al., 1999). Alternatively, subfunctionalization can occur on the protein level when one of the copies becomes specialized for only a certain aspect of the ancestral gene function (Hughes, 1999). An example for this are two paralogs of the RNA endonuclease gene in the archaea species *Sulfolobus solfataricus* (Tocchini-Valentini et al., 2005). The two genes encode different subunits of the orthologous RNA

endonuclease that is present in one copy in other archaea species, as for example, *Methanocaldococcus jannaschii*, and both of these subunits are required for enzymatic activity and cleavage of the pre-tRNA substrate. Another example for temporal gene subfunctionalization is the evolution of the  $\beta$ -globin cluster in humans. One gene from this cluster is expressed specifically in embryos, another in fetuses and another from birth onwards. In addition, each encodes a protein product with different oxygen binding affinity that is optimised for each developmental stage (Hurles, 2004). It has been proposed that genes with greater regulatory complexity are more likely to undergo subfunctionalization after duplication (Force et al., 1999), while the genes that are rapidly evolving, such as those involved in reproduction and immunity, are more likely to undergo neofunctionalization (Emes et al., 2003). In addition to the processes of neofunctionalization and subfunctionalization, gene duplication is sometimes a mechanism that ensures a higher level of gene expression (Zhang, 2003). In this scenario, it is beneficial to conserve the original function and it has been proposed that this is achieved either through frequent gene conversions and hence concerted evolution of the paralogues (Li, 1997) or through strong purifying selection against mutations that modify gene function (Nei et al., 2000). It is suggested that histones and ribosomal RNA genes have experienced several rounds of duplication because it was advantageous to increase expression of these essential genes in the cell (Hurles, 2004).

Gene duplications can also be a driving force for the evolution of novel domain arrangements. Firstly, point mutations in an already existing domain can create signatures of a novel domain with an original function (Weiner et al., 2006). Secondly, gene duplications can correlate with the creation of novel domain rearrangements (Vogel et al., 2005). Interestingly, duplicate genes in eukaryotes seem to have longer protein sequences and more functional domain than singleton genes (He and Zhang, 2005) Because of this, it was proposed that the majority of fixed duplicates undergoes sub- or neo-functionalization after duplication; complex genes are more likely to experience successful subfunctionalization and gene complexity can be regained after subsequent neofunctionalization (He and Zhang, 2005). An example for subfunctionalization on the level of domain arrangement is the one of the monkey king gene (mkg)

family in *Drosophila melanogaster* (Wang et al., 2004). Genes from the mkg family have originated recently as retroposed duplicates and due to complementary partial degradation evolved into fission genes that separately encode protein domains from a multidomain ancestor. Thus, gene duplication could result not only in the increase of a gene number, but also gene diversity. However, gene duplication is a slightly deleterious process and hence is more likely to become fixed in a population only when purifying selection is weak (Koonin, 2009). Since purifying selection is much weaker in smaller populations - such as the ones of higher eukaryotes, in contrast to bacteria - it has been suggested that there is no consistent tendency of evolution towards increased genomic complexity. Rather, that complexity is a non-adaptive consequence of evolution under low purifying selection (Koonin, 2009).

#### 1.2.4 Evolutionarily related proteins

A crucial step in studying protein evolution is to find related sequences and understand relationships between them. The concept of homology describes a relationship between genes or proteins that share a common evolutionary origin (Reeck et al., 1987). The terms orthology and paralogy have been introduced to extend the definition of homology; if the homology is the result of gene duplication the genes are defined as paralogous and if the homology is the result of speciation as orthologous (Fitch, 1970).

Databases that assign paralogous and orthologous proteins play a valuable role in finding homologous proteins and studying protein evolution. These databases either use pairwise protein comparisons to find the true orthologues, such as InParanoid (Berglund et al., 2008), use gene synteny to assist similarity as Ensembl Compara (Vilella et al., 2009), or build phylogenetic trees and base orthologue and paralogue assignments on them like TreeFam (Li et al., 2006).

### 1.3 Protein isoforms of the same gene

In the previous section, I addressed different means for the change of protein function during evolution. Point mutations, domain shuffling and gene duplications acted in concert to bring to expansion of the protein repertoire which was necessary for the emergence of more complex organisms. However, the number of genes in an organism shows a low correlation with the organismal complexity (Chothia et al., 2003). Therefore, a lot of attention has been drawn to the role of alternative splicing in the higher organisms (Flicek et al., 2010). Alternative splicing is quite abundant in the genomes of higher eukaryotes, with estimates that for example, there are on average four isoforms for every human gene (Melamud and Moul, 2009). Hence, this is a powerful mechanism for increasing protein diversity in an organism (illustrated in Figure 1.7). Similar to gene duplications, intron insertions are slightly deleterious, and it has been proposed that novel introns are also fixed only when the purifying selection is not strong (Koonin, 2009). Again, this implies that the resulting proteome diversity and organismal complexity were not actively selected for.

During splicing introns are removed from mRNA. Introns can vary substantially in size, but they maintain several conserved motifs, most prominently dinucleotides in their 5' and 3' ends - splice donors and splice acceptor sites. Since introns can be very long, it was suggested that splicing does not need to always operate by recognizing introns, but also by recognizing exons. Indeed, it has been reported that protein evolution is skewed in the vicinity intron-exon boundaries and shaped so that the nucleotide composition necessary for recognition and removal of introns is preserved (Parmley et al., 2007). Motifs that define intron positions in mRNA are recognized by components of the splicing machinery, which in turn recruit other components of the spliceosome – different snRNPs, which results in excision of an intron. Additional motifs inside introns and exons can determine alternative exon boundaries or exons that are included in the final product only in certain isoforms of a gene. Most likely, these events are regulated by additional splice factors. However, we still do not have a comprehensive knowledge of this process.

It has been noted that alternatively spliced exons in the human serine/arginine-rich (SR) family of splice regulators overlap with ultraconserved elements that are shared with mice (Lareau et al., 2007). Interestingly, it was shown that in every member of the human SR family, ultraconserved elements were recognized and alternatively spliced either as an alternative 'poison cassette exons' containing early in-frame stop codons, or as alternative introns in the 3' untranslated region (Lareau et al., 2007). These events target the resulting mRNAs for degradation by nonsense mediated mRNA decay (NMD). Since SR proteins direct splicing of their own products, this suggested that unproductive splicing is important for regulation of the entire SR family. Additionally, this also underlines the complexity of the alternative splicing regulation and implies an additional role for NMD. NMD is a surveillance mechanism that detects and degrades mRNAs with premature stop codons. Importantly, more than a third of reliably inferred alternative splicing events in humans result in mRNA isoforms with premature stop codons (Hillman et al., 2004). The fact that this phenomenon is so widespread indicates that NMD does not necessarily have a function to prevent protein mistranslation when errors occur, but could also be a regulatory mechanism that silences gene expression on posttranscriptional level.

Evolution of alternative splicing is tightly linked to protein evolution. Interestingly, one of the mechanisms for generating new cassette exons – exons that are excluded or included in a processed mRNA with their whole length – is exon shuffling (Kondrashov and Koonin, 2003; Letunic et al., 2002). By this means, either a new exon is inserted into a gene, or an existing exon is duplicated within a gene. Alternative cassette exons can also emerge through exonization of intronic sequences (Wang et al., 2005). Close to 5% of human genes contain motifs of transposable elements in their coding regions, such as of Alu elements (Sorek et al., 2002). Importantly, newly inserted exons often have a low inclusion level, thus the ancestral mRNA remains the main gene product (Mendes Soares and Valcarcel, 2006). In line with this, alternative cassette exons with a high inclusion level are usually conserved between human and mouse, which is not the case for those with a low inclusion level (Modrek and Lee, 2003). In addition to this, alternatively spliced exons can also originate from the constitutive



ancestral exons - exons present in all splice isoforms of a gene - through creation of novel splice sites (Lev-Maor et al., 2007).

New sequencing technologies are making the studies of alternative splicing more comprehensive (Pan et al., 2008) and will surely have a great impact on the understanding of this process, but potentially also on disease treatment. By now, alternative splicing has been implicated in a number of human genetic diseases; in particular different neurodegenerative disorders and cancer (Lukong et al., 2008). At this time, therapeutic strategies that target splicing defects look promising. A number of these are underway and some, such as agents that target splicing factors or isoform-specific drugs are already in use (Garcia-Blanco et al., 2004). An example for the former is an inhibitor of the Clk1/Sly kinase, which phosphorylates SR proteins, and for the latter is phenacetin, a nonsteroidal anti-inflammatory drug that has a different inhibitory effect on the activity of different isoforms of the COX enzyme. However, the role of alternative splicing in disease development is most probably still underappreciated. We do not have a knowledge of all regulatory signals for gene splicing and even synonymous mutations that are usually discarded as disease causing can affect splicing and disrupt the protein (Caceres and Kornblihtt, 2002). Moreover, if the mutated gene interacts with a number of molecular partners then the effects of the observed mutation should be viewed in the context of the whole molecular network (Schadt, 2009).

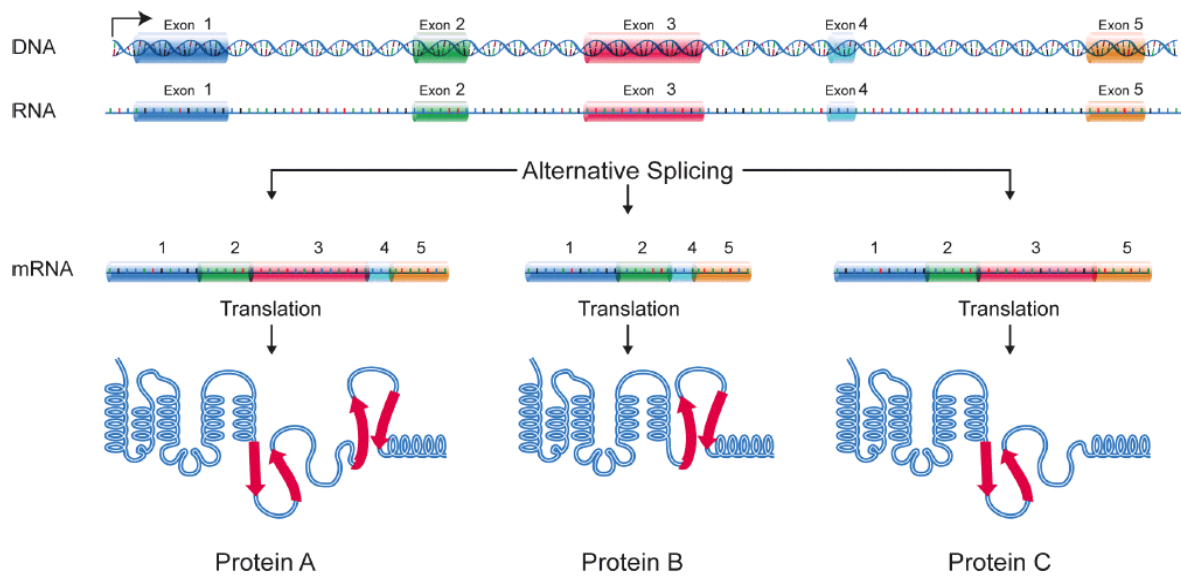


Figure 1.7: Alternative splicing increases the diversity of proteome. Alternative inclusion of exons 3 and 4 in this example can change the structure and function of the resulting protein products. Figure is taken from: [http://upload.wikimedia.org/wikipedia/commons/0/0a/DNA\\_alternative\\_splicing.gif](http://upload.wikimedia.org/wikipedia/commons/0/0a/DNA_alternative_splicing.gif)

## 1.4 Outline of the thesis

The remaining chapters of this thesis consist of three separate investigations. I first analyse general trends in the evolution of protein domain architectures. This analysis lays a foundation for the work in the following chapter where I focus on the smaller set of confident domain gain events and investigate molecular mechanisms that underlined these domain insertions. In the final results chapter, I analyse characteristics of protein regions that undergo tissue-specific alternative splicing. Thus, the overall aim of this thesis is to address changes in the architecture of protein functional elements on different levels.

Parts of the results described in Chapters 2 and 3 have been published (Buljan and Bateman, 2009; Buljan et al., 2010). Work in Chapter 4 is in preparation for submission at the time when the thesis is submitted.

## 1.5 Bibliography

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Altschul, S.F., and Koonin, E.V. (1998). Iterated profile searches with PSI-BLAST-- a tool for discovery in protein databases. *Trends Biochem Sci* 23, 444-447.
- Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36, D419-425.
- Arguello, J.R., Chen, Y., Yang, S., Wang, W., and Long, M. (2006). Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. *PLoS Genet* 2, e77.
- Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., et al. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 31, 400-402.
- 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.
- Bartlett, G.J., Borkakoti, N., and Thornton, J.M. (2003). Catalysing new reactions during evolution: economy of residues and mechanism. *J Mol Biol* 331, 829-860.
- 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 Res* 30, 276-280.
- Berglund, A.C., Sjolund, E., Ostlund, G., and Sonnhammer, E.L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 36, D263-266.
- Betran, E., Thornton, K., and Long, M. (2002). Retroposed new genes out of the X in *Drosophila*. *Genome Res* 12, 1854-1859.
- Bjorklund, A.K., Ekman, D., and Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Comput Biol* 2, e114.

- 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 Res* 33, D212-215.
- Buljan, M., and Bateman, A. (2009). The evolution of protein domain families. *Biochem Soc Trans* 37, 751-755.
- Buljan, M., Frankish, A., and Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol* 11, R74.
- Burki, F., and Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet* 36, 1061-1063.
- Caceres, J.F., and Kornblihtt, A.R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* 18, 186-193.
- Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.Y., Romero, P., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2006). Rational drug design via intrinsically disordered protein. *Trends Biotechnol* 24, 435-442.
- Chothia, C., and Gough, J. (2009). Genomic and structural aspects of protein evolution. *Biochem J* 419, 15-28.
- Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science* 300, 1701-1703.
- Ciccarelli, F.D., von Mering, C., Suyama, M., Harrington, E.D., Izaurralde, E., and Bork, P. (2005). Complex genomic rearrangements lead to novel primate gene function. *Genome Res* 15, 343-351.
- Cohen, P. (2001). The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur J Biochem* 268, 5001-5010.
- Coin, L. (2008). Protein Domains: New methods for detection and evolutionary analysis. In Wellcome Trust Sanger Institute (Cambridge, University of Cambridge).
- Conant, G.C., and Wagner, A. (2005). The rarity of gene shuffling in conserved genes. *Genome Biol* 6, R50.
- Deininger, P.L., Moran, J.V., Batzer, M.A., and Kazazian, H.H., Jr. (2003). Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* 13, 651-658.

- Denning, D.P., Patel, S.S., Uversky, V., Fink, A.L., and Rexach, M. (2003). Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc Natl Acad Sci U S A* 100, 2450-2455.
- Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I., and Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 5, 2985-2995.
- 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. *J Mol Biol* 347, 827-839.
- Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. (2001). Intrinsically disordered protein. *J Mol Graph Model* 19, 26-59.
- Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6, 197-208.
- Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755-763.
- Ekman, D., Bjorklund, A.K., and Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol* 372, 1337-1348.
- 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. *J Mol Biol* 348, 231-243.
- Emes, R.D., Goodstadt, L., Winter, E.E., and Ponting, C.P. (2003). Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum Mol Genet* 12, 701-709.
- Feng, D.F., Cho, G., and Doolittle, R.F. (1997). Determining divergence times with a protein clock: update and reevaluation. *Proc Natl Acad Sci U S A* 94, 13028-13033.
- Fernandez-Fuentes, N., Dybas, J.M., and Fiser, A. (2010). Structural characteristics of novel protein folds. *PLoS Comput Biol* 6, e1000750.
- Finn, R.D., Mistry, J., Tate, J., Coghill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. *Nucleic Acids Res* 38, D211-222.
- Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* 19, 99-113.

- Flicek, P., Aken, B.L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., et al. (2010). Ensembl's 10th year. *Nucleic Acids Res* 38, D557-562.
- Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531-1545.
- Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23, 950-956.
- Gahmberg, C.G., and Tolvanen, M. (1996). Why mammalian cell surface proteins are glycoproteins. *Trends Biochem Sci* 21, 308-311.
- Garcia-Blanco, M.A., Baraniak, A.P., and Lasda, E.L. (2004). Alternative splicing in disease and therapy. *Nat Biotechnol* 22, 535-546.
- Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A.K. (1998). Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. *Genome Inform Ser Workshop Genome Inform* 9, 201-213.
- Goldstein, R.A. (2008). The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol* 18, 170-177.
- Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., et al. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35, D291-297.
- Gribskov, M., McLachlan, A.D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* 84, 4355-4358.
- Gsponer, J., and Babu, M.M. (2009). The rules of disorder or why disorder rules. *Prog Biophys Mol Biol* 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* 322, 1365-1368.
- Gu, X., Wang, Y., and Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet* 31, 205-209.

- H Lodish, A.B., S L Zipursky, P Matsudaira, D Baltimore, and J Darnell (2000). *Molecular Cell Biology*, 4th edition edn (New York, W. H. Freeman).
- Haber, J.E. (2000). Partners and pathways repairing a double-strand break. *Trends Genet* 16, 259-264.
- Haft, D.H., Selengut, J.D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res* 31, 371-373.
- Haines, N., and Irvine, K.D. (2003). Glycosylation regulates Notch signalling. *Nat Rev Mol Cell Biol* 4, 786-797.
- Han, J.H., Batey, S., Nickson, A.A., Teichmann, S.A., and Clarke, J. (2007). The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol* 8, 319-330.
- Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., and Iakoucheva, L.M. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* 2, e100.
- He, X., and Zhang, J. (2005). Gene complexity and gene duplicability. *Curr Biol* 15, 1016-1021.
- Heger, A., Wilton, C.A., Sivakumar, A., and Holm, L. (2005). ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res* 33, D188-191.
- Henikoff, J.G., Pietrokovski, S., McCallum, C.M., and Henikoff, S. (2000). Blocks-based methods for detecting protein homology. *Electrophoresis* 21, 1700-1706.
- Hillman, R.T., Green, R.E., and Brenner, S.E. (2004). An unappreciated role for RNA surveillance. *Genome Biol* 5, R8.
- Holm, L., and Sander, C. (1994). Parser for protein folding units. *Proteins* 19, 256-268.
- Hughes, A.L. (1999). *Adaptive evolution of genes and genome* (New York, Oxford University Press).
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M., and Sigrist, C.J. (2006). The PROSITE database. *Nucleic Acids Res* 34, D227-230.



- Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res* 37, D211-215.
- Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. *PLoS Biol* 2, E206.
- Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., and Dunker, A.K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323, 573-584.
- Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32, 1037-1049.
- Jacob, F. (1977). Evolution and tinkering. *Science* 196, 1161-1166.
- Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. (2001). Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413, 514-519.
- Kaessmann, H. (2009). Genetics. More than just a copy. *Science* 325, 958-959.
- Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* 10, 19-31.
- Kall, L., Krogh, A., and Sonnhammer, E.L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338, 1027-1036.
- Knowles, D.G., and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Res* 19, 1752-1759.
- Kondrashov, F.A., and Koonin, E.V. (2003). Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet* 19, 115-119.
- Koonin, E.V. (2009). Darwinian evolution in the light of genomics. *Nucleic Acids Res* 37, 1011-1034.
- Landry, C.R., Levy, E.D., and Michnick, S.W. (2009). Weak functional constraints on phosphoproteomes. *Trends Genet* 25, 193-197.
- Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* 446, 926-929.

- Letunic, I., Copley, R.R., and Bork, P. (2002). Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet* 11, 1561-1567.
- Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34, D257-260.
- Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., Leibman-Barak, S., Pupko, T., and Ast, G. (2007). The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet* 3, e203.
- 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 Res* 34, D572-580.
- Li, W.H. (1997). *Molecular Evolution* (Sunderland Massachusetts, Sinauer Associates, Inc.).
- Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31, 3701-3708.
- Long, M. (2001). Evolution of novel genes. *Curr Opin Genet Dev* 11, 673-680.
- Ludin, B., Ashbridge, K., Funfschilling, U., and Matus, A. (1996). Functional analysis of the MAP2 repeat domain. *J Cell Sci* 109 ( Pt 1), 91-99.
- Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends Genet* 24, 416-425.
- Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A* 107, 9270-9274.
- Mann, M., and Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. *Nat Biotechnol* 21, 255-261.
- Marsh, J.A., and Teichmann, S.A. (2010). How do proteins gain new domains? *Genome Biol* 11, 126.
- Melamud, E., and Moul, J. (2009). Structural implication of splicing stochasticity. *Nucleic Acids Res* 37, 4862-4872.
- Mendes Soares, L.M., and Valcarcel, J. (2006). The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J* 25, 923-931.

- Modrek, B., and Lee, C.J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* 34, 177-180.
- Moore, A.D., Bjorklund, A.K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends Biochem Sci* 33, 444-451.
- Neduva, V., and Russell, R.B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Lett* 579, 3342-3345.
- Nei, M., Rogozin, I.B., and Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A* 97, 10866-10871.
- Neurath, H., and Walsh, K.A. (1976). Role of proteolytic enzymes in biological regulation (a review). *Proc Natl Acad Sci U S A* 73, 3825-3832.
- Ohno, S. (1970). *Evolution by gene duplication* (Berlin, Springer-Verlag).
- Pal, C., Papp, B., and Lercher, M.J. (2006). An integrated view of protein evolution. *Nat Rev Genet* 7, 337-348.
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40, 1413-1415.
- Parmley, J.L., Urrutia, A.O., Potrzebowski, L., Kaessmann, H., and Hurst, L.D. (2007). Splicing and the evolution of proteins in mammals. *PLoS Biol* 5, e14.
- Pasek, S., Risler, J.L., and Brezellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* 22, 1418-1423.
- Patthy, L. (1996). Exon shuffling and other ways of module exchange. *Matrix Biol* 15, 301-310; discussion 311-302.
- Pawson, T. (2003). Organization of cell-regulatory systems through modular-protein-interaction domains. *Philos Transact A Math Phys Eng Sci* 361, 1251-1262.
- Pils, B., and Schultz, J. (2004). Inactive enzyme-homologues find new function in regulatory processes. *J Mol Biol* 340, 399-404.
- Reeck, G.R., de Haen, C., Teller, D.C., Doolittle, R.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., et al. (1987).

- "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* 50, 667.
- Resh, M.D. (1999). Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins. *Biochim Biophys Acta* 1451, 1-16.
- Rosenberg, H.F. (1995). Recombinant human eosinophil cationic protein. Ribonuclease activity is not essential for cytotoxicity. *J Biol Chem* 270, 7876-7881.
- Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., and Liberles, D.A. (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol* 308, 58-73.
- Rueter, S.M., Dawson, T.R., and Emeson, R.B. (1999). Regulation of alternative splicing by RNA editing. *Nature* 399, 75-80.
- Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218-223.
- Shimizu, K., and Toh, H. (2009). Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. *J Mol Biol* 392, 1253-1265.
- Siepel, A. (2009). Darwinian alchemy: Human genes from noncoding DNA. *Genome Res* 19, 1693-1695.
- Sorek, R., Ast, G., and Graur, D. (2002). Alu-containing exons are alternatively spliced. *Genome Res* 12, 1060-1067.
- Tocchini-Valentini, G.D., Fruscoloni, P., and Tocchini-Valentini, G.P. (2005). Structure, function, and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. *Proc Natl Acad Sci U S A* 102, 8933-8938.
- Tokuriki, N., and Tawfik, D.S. (2009). Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* 459, 668-673.
- Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579, 3346-3354.
- Tompa, P., and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *FASEB J* 18, 1169-1175.

- Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19, 327-335.
- Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. (2004). Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14, 208-216.
- Vogel, C., Teichmann, S.A., and Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *J Mol Biol* 346, 355-365.
- Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., et al. (2005). DisProt: a database of protein disorder. *Bioinformatics* 21, 137-140.
- Walsh, C.T. (2006). *Posttranslational Modification of Proteins* (Englewood, Colorado, Roberts and Company Publishers).
- Wang, W., Yu, H., and Long, M. (2004). Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat Genet* 36, 523-527.
- Wang, W., Zheng, H., Yang, S., Yu, H., Li, J., Jiang, H., Su, J., Yang, L., Zhang, J., McDermott, J., et al. (2005). Origin and evolution of new exons in rodents. *Genome Res* 15, 1258-1264.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337, 635-645.
- Weiner, J., 3rd, Beaussart, F., and Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *FEBS J* 273, 2037-2047.
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* 37, D380-386.
- Wojtowicz, W.M., Flanagan, J.J., Millard, S.S., Zipursky, S.L., and Clemens, J.C. (2004). Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell* 118, 619-633.
- Wootton, J.C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 18, 269-285.

- Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293, 321-331.
- Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X., and Orengo, C. (2008). Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res* 36, D414-418.
- Yokoyama, S., and Yokoyama, R. (1989). Molecular evolution of human visual pigment genes. *Mol Biol Evol* 6, 186-197.
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18, 292-298.
- Zhang, J., Rosenberg, H.F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* 95, 3708-3713.
- Zhang, L., Ma, B., Wang, L., and Xu, Y. (2003). Greedy method for inferring tandem duplication history. *Bioinformatics* 19, 1497-1504.
- Zheng, D., Zhang, Z., Harrison, P.M., Karro, J., Carriero, N., and Gerstein, M. (2005). Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J Mol Biol* 349, 27-45.

## 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).

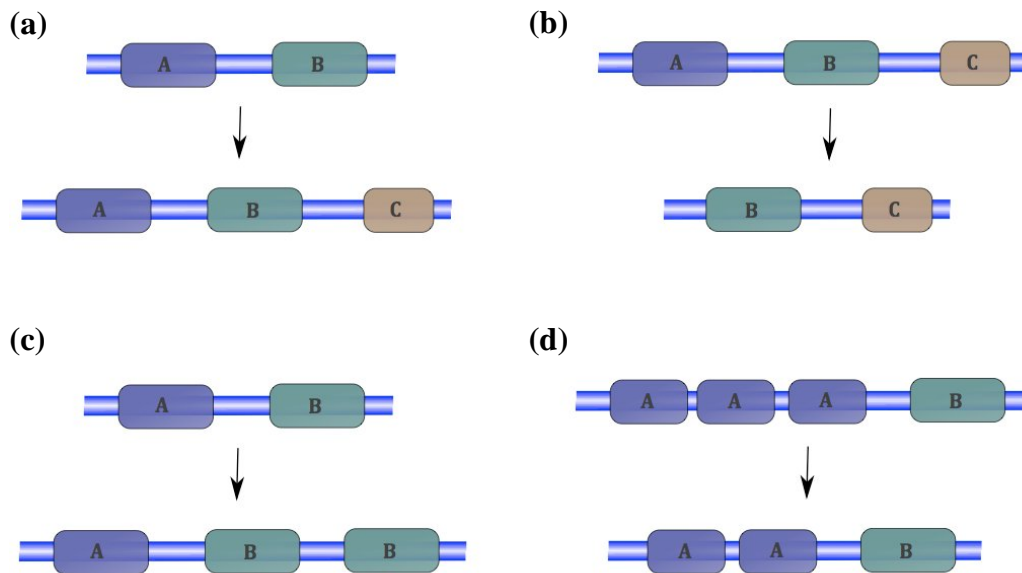


Figure 2.1 Illustration of domain gains and losses. Figure (a) illustrates gain of a novel domain and figure (b) loss of a domain, which was present in one copy in the ancestral protein. Figure (c) illustrates a domain gain, which leads to a domain repeat and figure (d) loss of a domain from a repeat.

## 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 Phylogenetic 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 homologous 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 coverage 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%

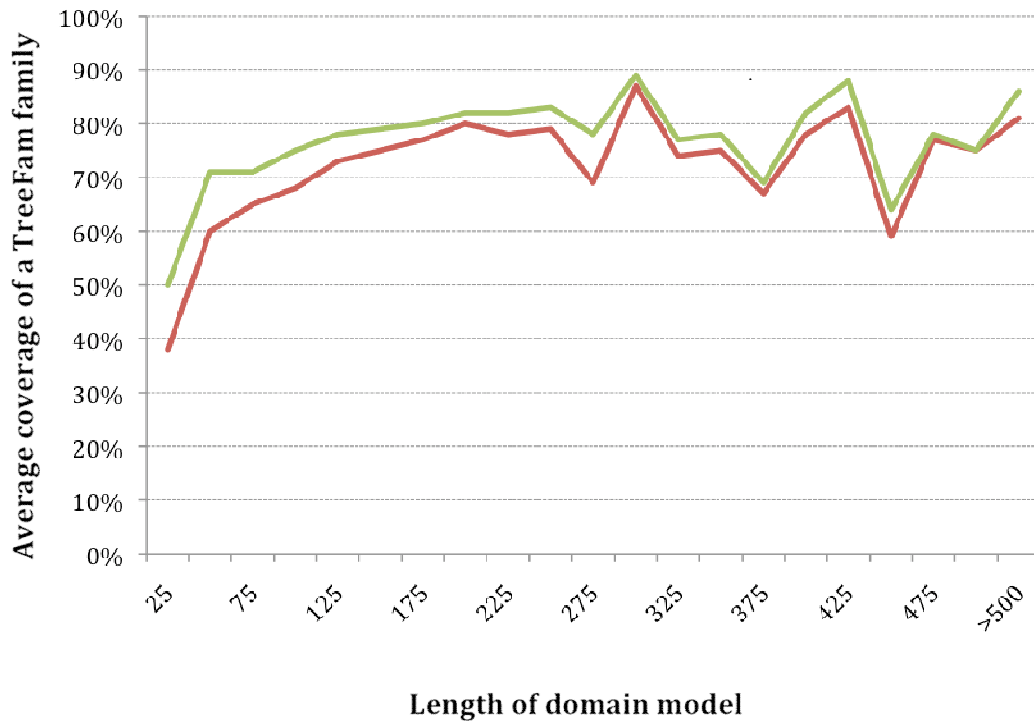


Figure 2.2: Average coverage of TreeFam gene families with Pfam domains of different lengths. Consistency of domain annotations between the members of the same TreeFam-A family represents the quality of domain annotations. Model lengths are grouped in bin categories of 25 amino acids, and all domains 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;  $2.2 \times 10^{-16}$  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.

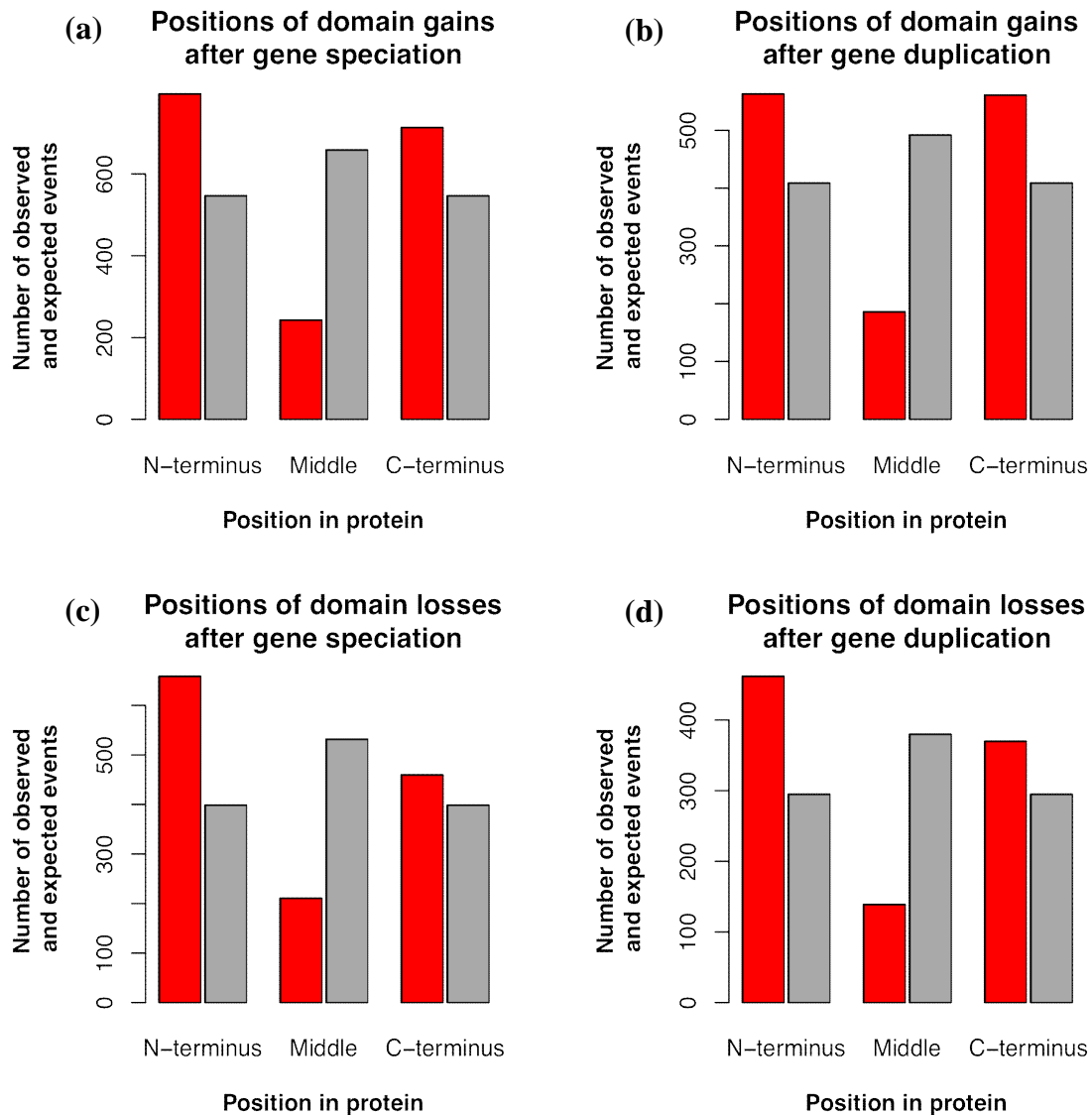


Figure 2.3: Positions of changes in proteins. Positions in proteins where gains (a and b) and losses (c and d) of single copy domains have been observed after gene speciation (a and c) and duplication (b and d) are shown. Observed and expected numbers of events are presented as red and grey columns, 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 x10 <sup>-16</sup>
		Middle	243	659	
		C-terminus	714	547	
	Domain loss	N-terminus	659	399	P<2.2 x10 <sup>-16</sup>
		Middle	211	532	
		C-terminus	460	399	
Gene duplication	Domain gain	N-terminus	563	409	P<2.2 x10 <sup>-16</sup>
		Middle	186	492	
		C-terminus	561	409	
	Domain loss	N-terminus	462	295	P<2.2 x10 <sup>-16</sup>
		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.



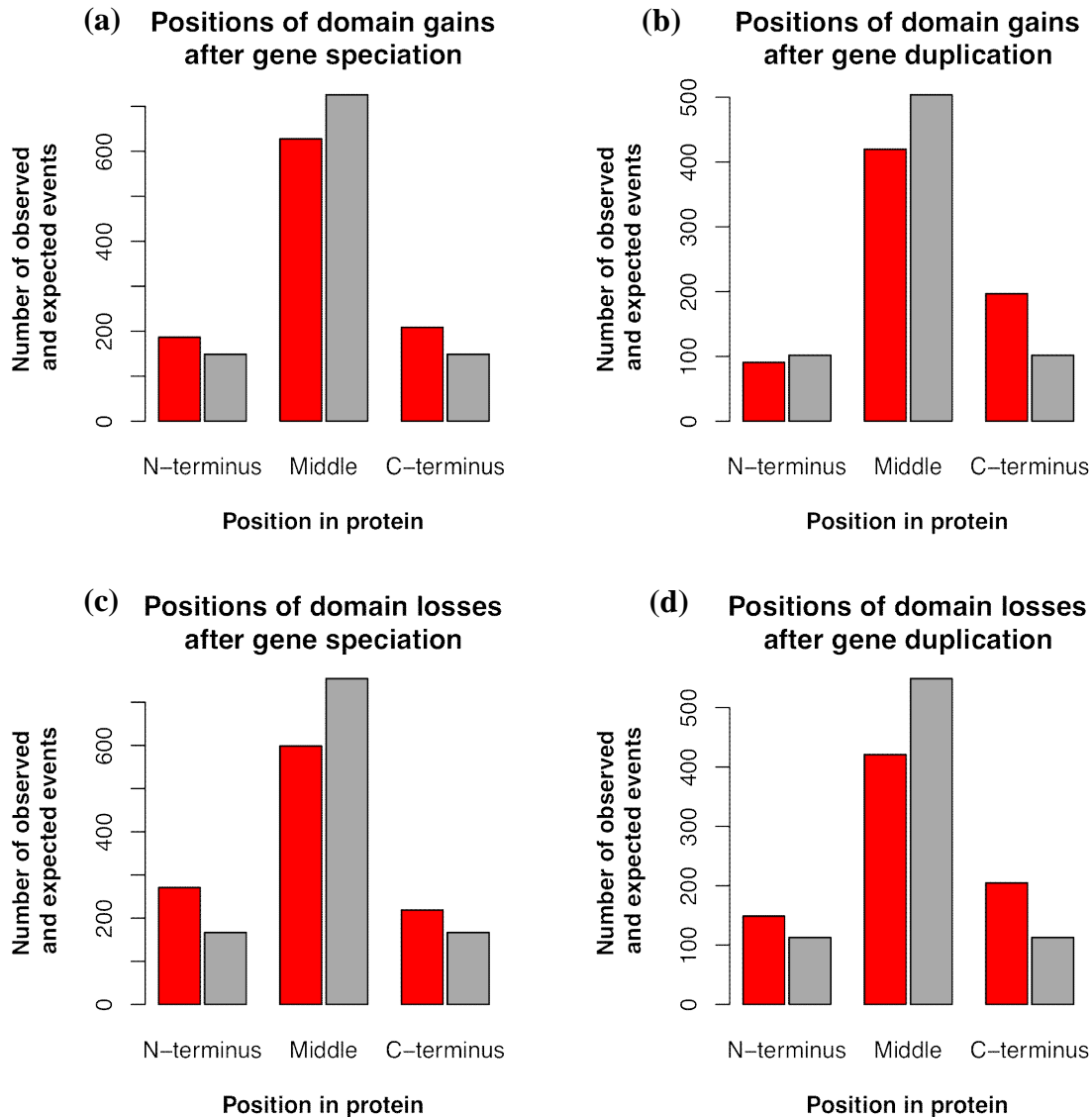


Figure 2.4: Positions of gains and losses of domains in repeats. Positions in proteins where gains (a and b) and losses (c and d) of domains in repeats have been observed after gene speciation (a and c) and duplication (b and d). Observed and expected numbers of events are presented as red and grey columns, 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 x10 <sup>-11</sup>
		Middle	628	726	
		C-terminus	209	149	
	Domain loss	N-terminus	271	167	P<2.2 x10 <sup>-16</sup>
		Middle	599	755	
		C-terminus	219	167	
Gene duplication	Domain gain	N-terminus	91	102	P<2.2 x10 <sup>-16</sup>
		Middle	420	504	
		C-terminus	197	102	
	Domain loss	N-terminus	149	113	P<2.2 x10 <sup>-16</sup>
		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 than 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 than 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 superfamily	Extracellular processes
85	PF00096	Zinc finger, C2H2 type	Regulation: DNA-binding
76	CL0011	Immunoglobulin 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 than 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.

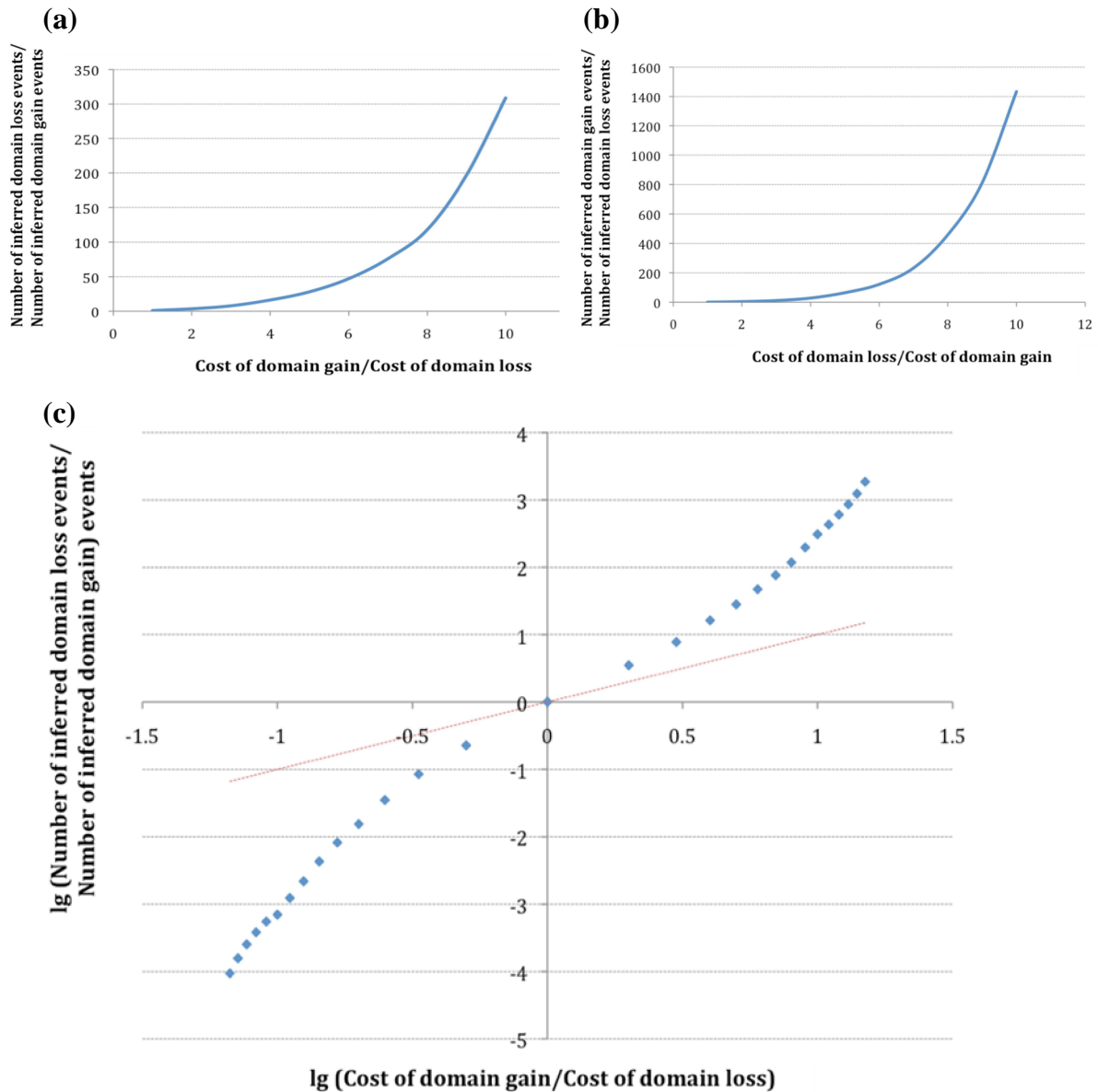


Figure 2.5: The ratio of inferred domain gain and loss events strongly depends on the assumed cost of these events. (a) The ratio of inferred domain loss and gain events exponentially depends on the ratio of increasing assumed cost for domain gain and loss event. The higher the cost of an event, the smaller is the likelihood of observing the event. (b) Similarly to (a), increasing the cost of domain loss results in an exponential 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 represents 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 inferred ratio 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.

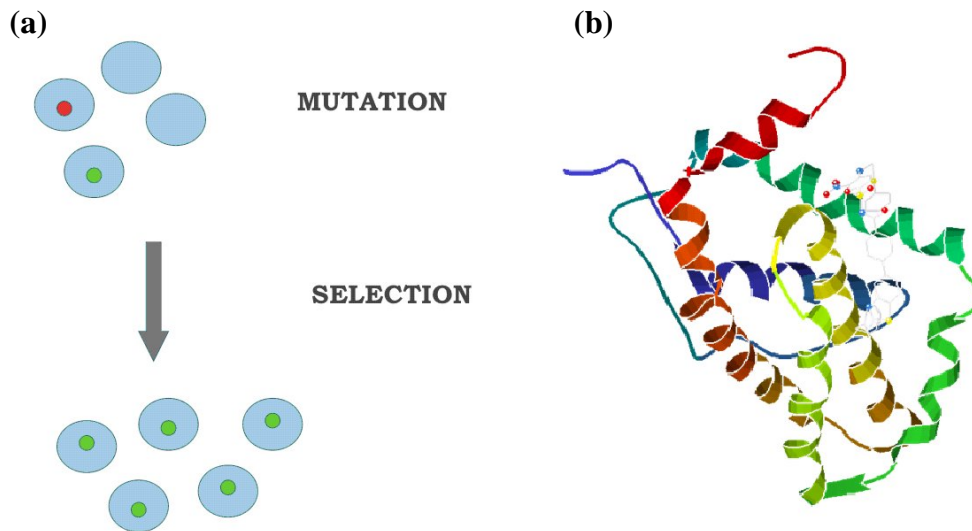


Figure 2.6: The evolution of domain architectures is determined by molecular mechanisms that cause the changes as well as subsequent selection. (a) Different molecular mechanisms can cause changes in domain architecture, but only some of the created architectures survive the subsequent evolutionary selection. 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., Kyripides, 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.

## 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.

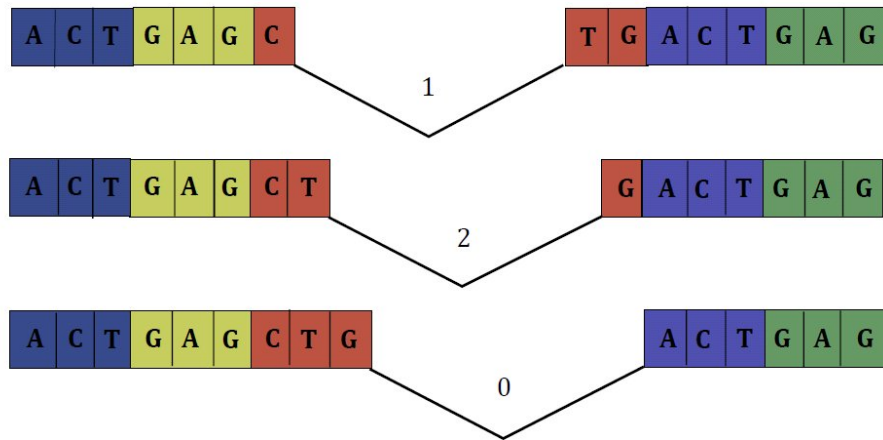


Figure 3.1. Illustration of intron phases. Phase of an intron is defined by the breakpoint in the codon adjacent to the intron.

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.

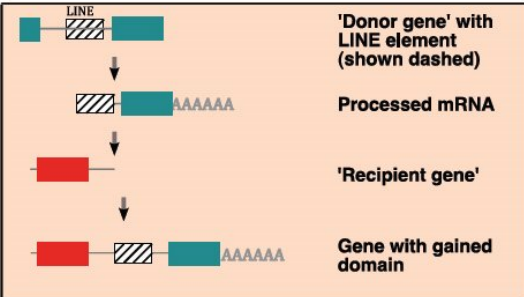
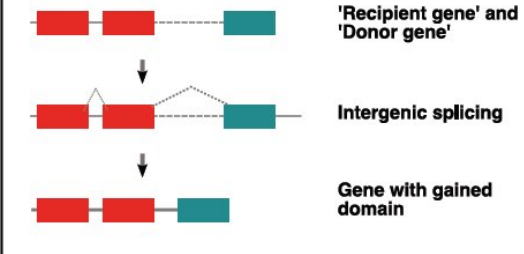

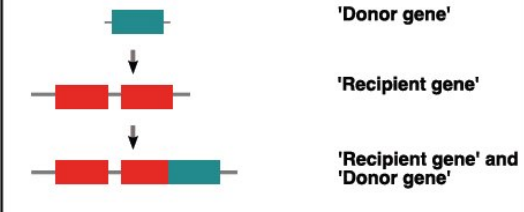
Result of domain gain	Possible causative mechanism	Position of gain	Number of exons gained
 <p>'Donor gene' with LINE element (shown dashed)</p> <p>Processed mRNA</p> <p>'Recipient gene'</p> <p>Gene with gained domain</p>	Retroposition.	Anywhere in a protein.	Only one, since the intermediate step is 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.

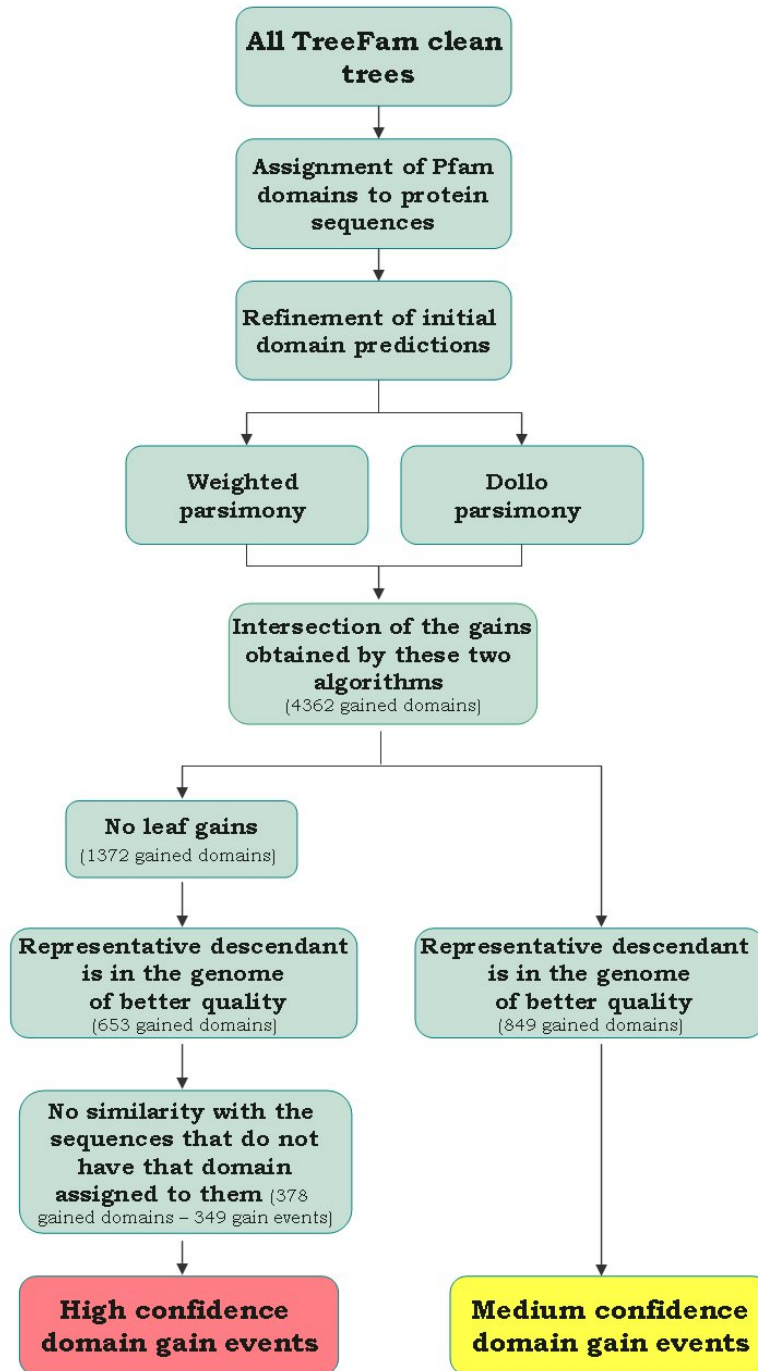


Figure 3.3: Flowchart of methods for obtaining sets of high and medium confidence domain gain. The numbers of gained domains I was left with after each filtering step are noted. In some cases more domains were gained at the same time; hence the number of gain events that we looked at for the high confidence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

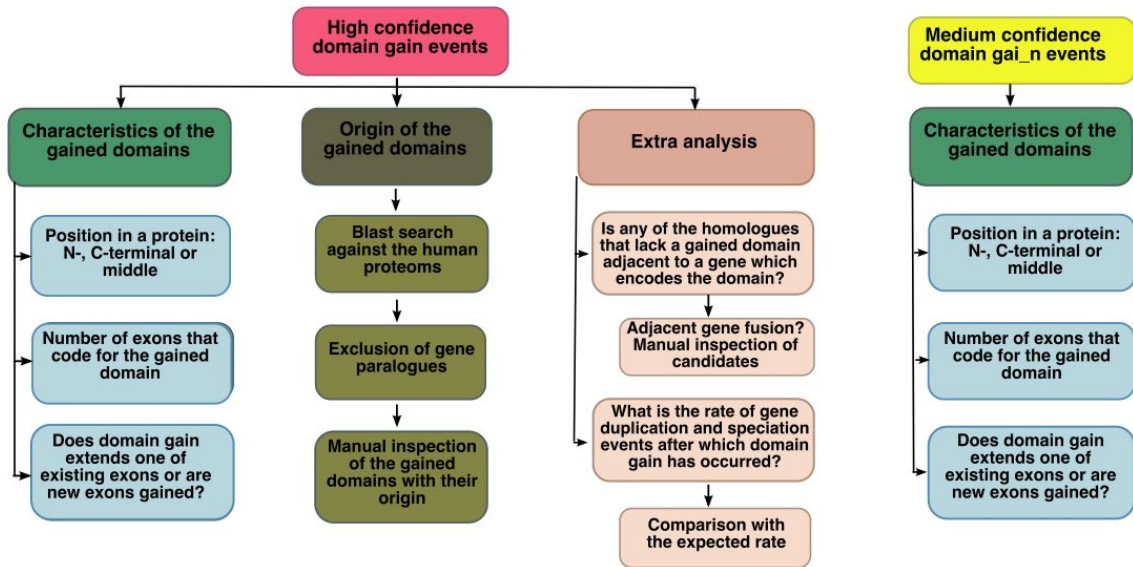


Figure 3.4: Flowchart of analysis for the sets of high and medium confidence domain gain events. For the set of high confidence domain gain events, I looked at characteristics of the gained domains, their potential origin and other trends that could imply potential causal mechanism. For the set of medium 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.

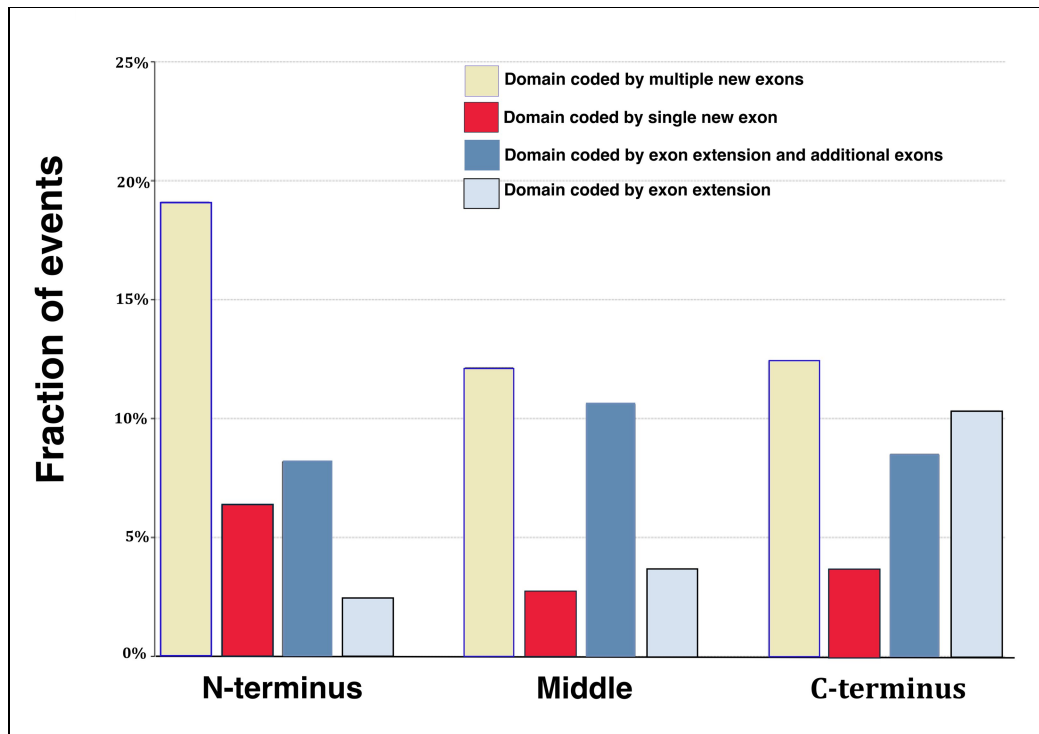


Figure 3.5: Distribution of domain gain events in the high confidence set of domain gains according to the position of domain insertion and number of exons gained. Gains at N- and C- termini and in the middle of proteins are shown separately. The first column in each group shows the fraction of gains 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.

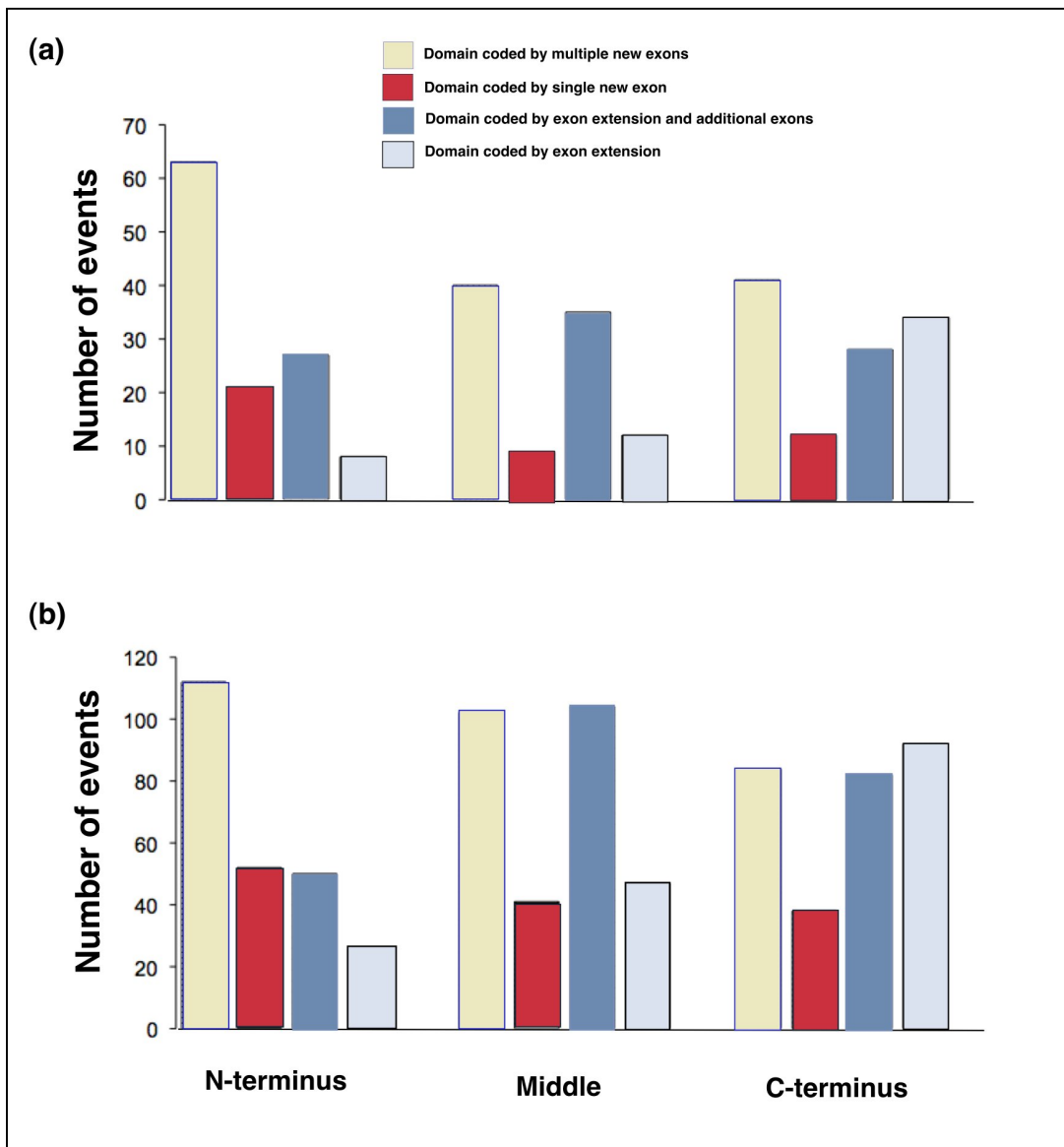


Figure 3.6: Distribution of domain gain events according to the position of domain insertion and number of exons gained in the set of high confidence domain gains and in the set of medium confidence domain gains. Distribution of characteristics of domains from the high confidence set of domain gains (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.

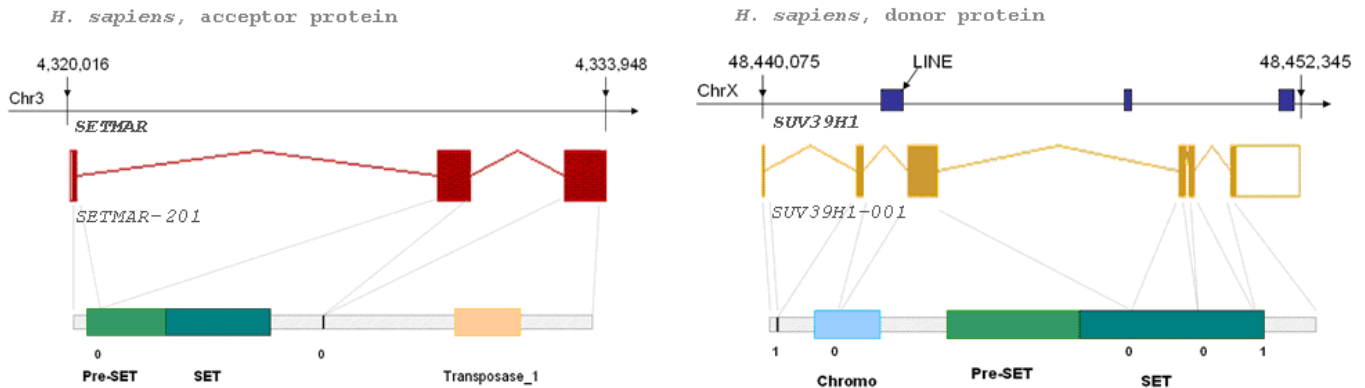


Figure 3.7: Retroposition as a causing mechanism for domain gain.

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.

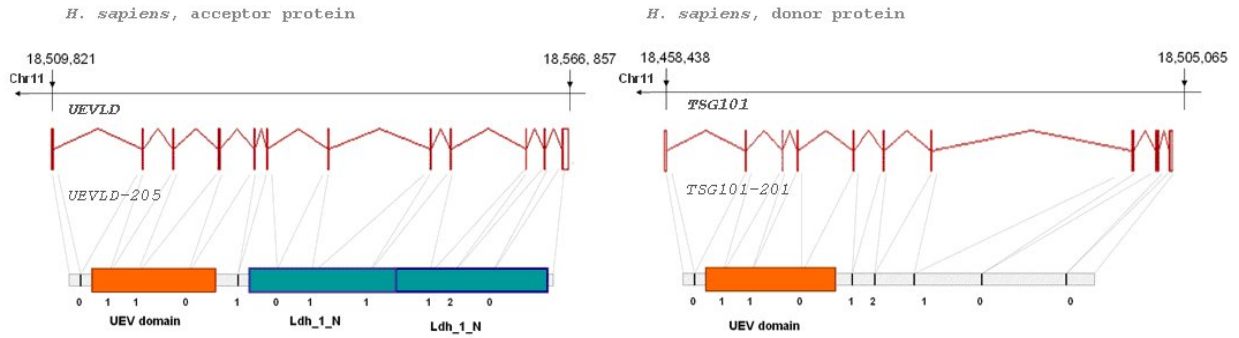
The search for the 'donor gene' of the gained domains identified the possible origin of the domain for 60% of domains coded by new terminal exons. This implies that duplication of a donor domain has frequently provided the material for subsequent exon joining and new exon combinations. An illustration 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

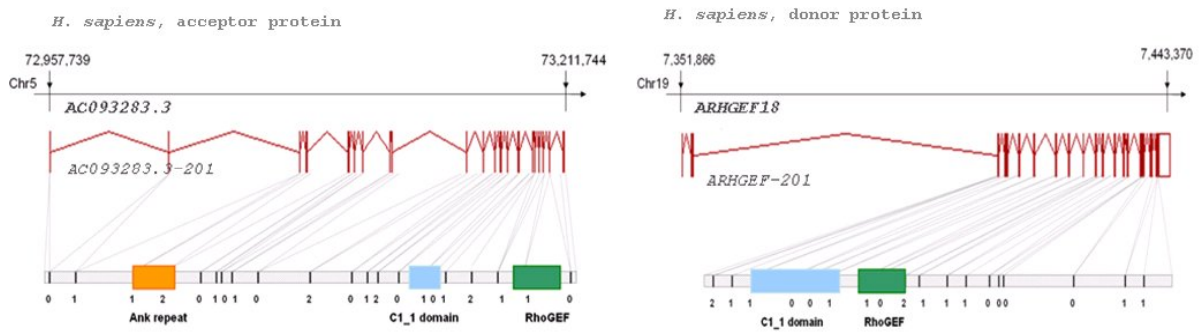
lineage did not find LINE elements that preceded a gained or 'donor' domain and hence did not lend support for retroposition as a causative mechanism (described above). With regard to other categories of domain gain events in Figure 3.5., because of the strict criteria I used to call a gained domain 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)

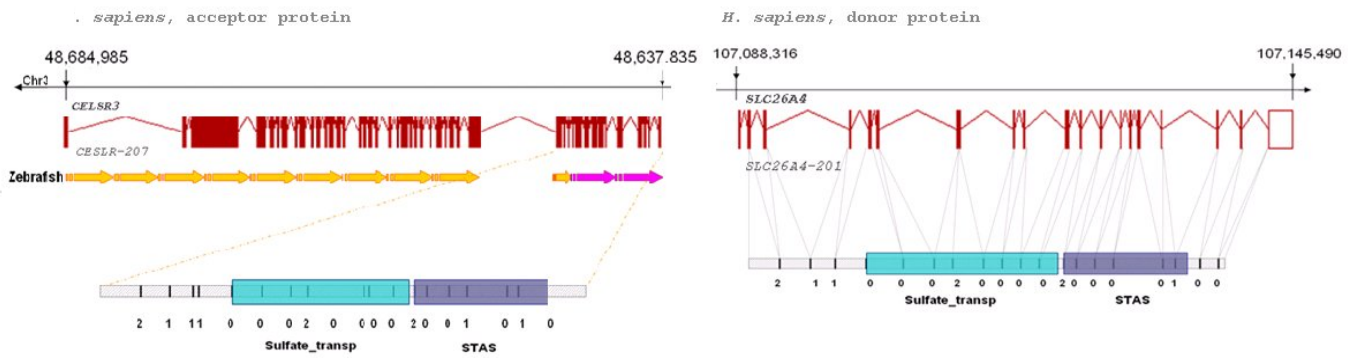


Figure 3.8: Examples for domain gains by joining of exons from two ancestral genes. A representative protein for a domain gain is always shown on the left and a protein which is a potential origin of the gained domain is shown on the right. (a) An example of a domain gain by gene duplication followed 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.

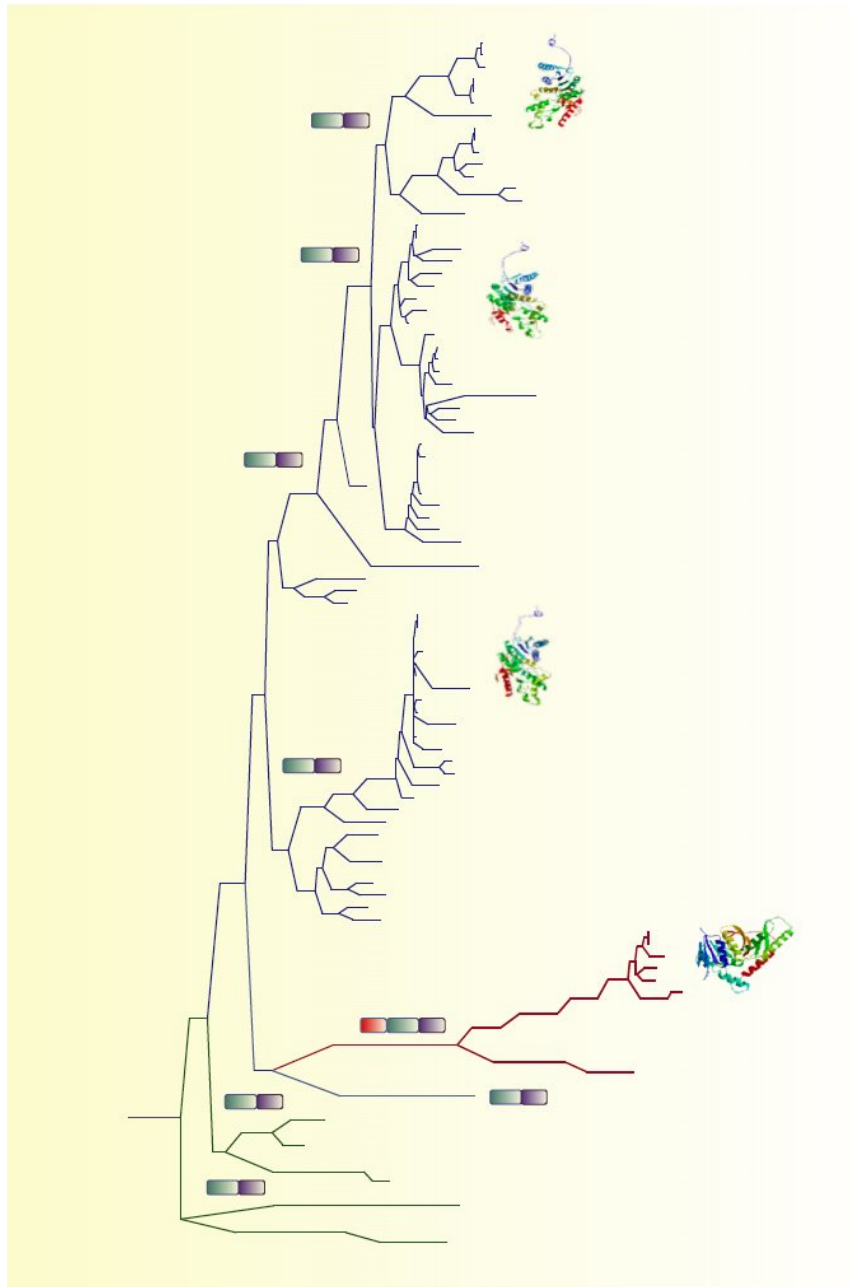


Figure 3.9: Gain of the UEV domain in the TreeFam family TF334740. Structure of a representative gene that was extended with the UEV domain is shown in Figure 3.8a. Here, the evolutionary tree of lactate dehydrogenase genes is shown. Vertebrate genes in the tree – the red coloured branch – have gained 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.





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.

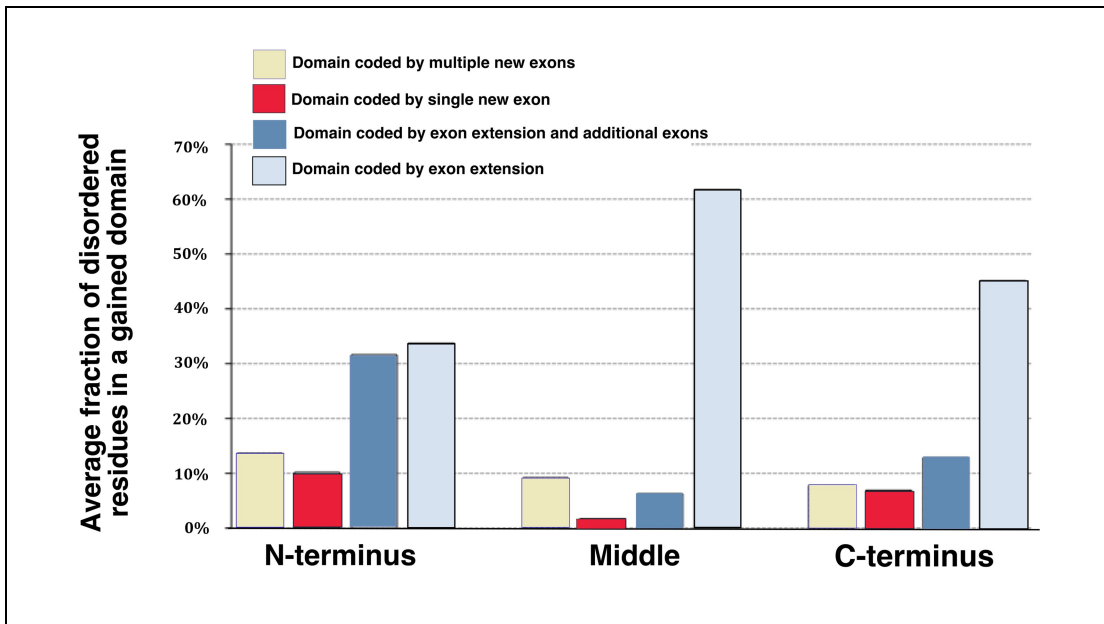


Figure 3.11: Distribution of disordered residues in the gained domains according to the position of domain insertion and number of exons gained. This graph shows the percentage of disordered residues in each category of domain gains. The number of events in each category can be seen in Figure 3.5.

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<sup>+</sup> 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.

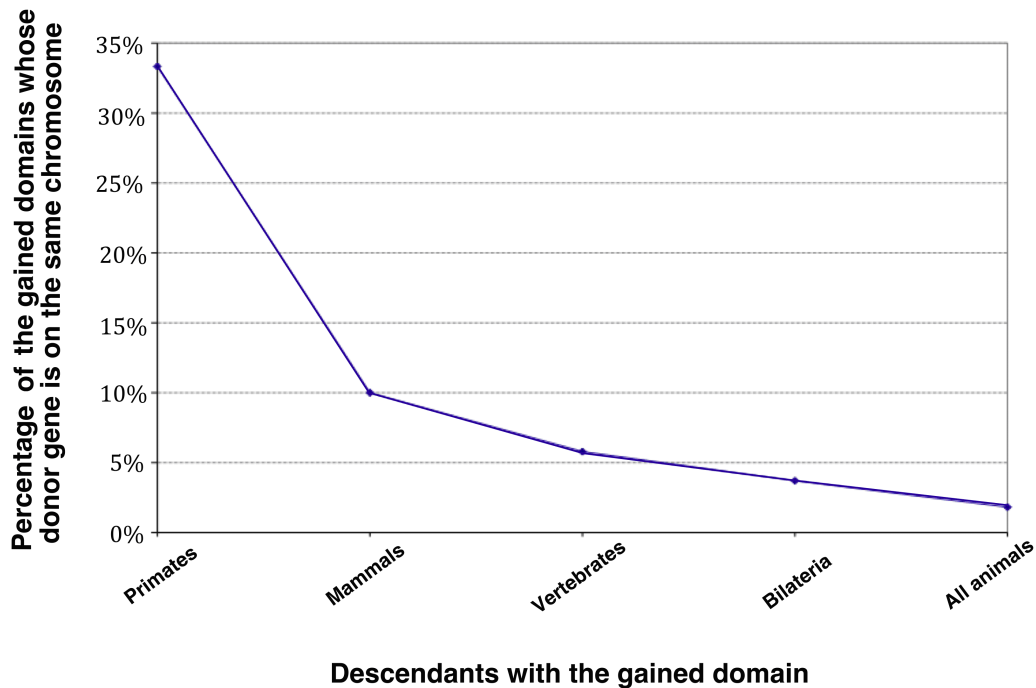


Figure 3.12: Chromosomal position of the ‘donor gene’ and the relative age of the gain event. The graph is showing the fraction of events for which the ‘donor gene’ of the gained domain is identified, and is on the same chromosome as the gene with the gained domain, with respect to the relative age of 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 Bilateralia 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 <sup>-6</sup>	3.1 x10 <sup>-4</sup>
	CC	0005924	cell-substrate adherens junction	4.3 x10 <sup>-5</sup>	5.8 x10 <sup>-3</sup>
	CC	0030055	cell-substrate junction	6.3 x10 <sup>-5</sup>	5.8 x10 <sup>-3</sup>
	CC	0005925	focal adhesion	2.3 x10 <sup>-4</sup>	1.3 x10 <sup>-2</sup>
	CC	0005912	adherens junction	5.9 x10 <sup>-4</sup>	2.7 x10 <sup>-2</sup>
	CC	0070161	anchoring junction	1.2 x10 <sup>-3</sup>	4.5 x10 <sup>-2</sup>
Annotation Cluster 2	BP	0006793	phosphorus metabolic process	5.4 x10 <sup>-6</sup>	9.2x10 <sup>-3</sup>
	BP	0006796	phosphate metabolic process	5.4 x10 <sup>-6</sup>	9.2x10 <sup>-3</sup>
	MF	0030554	adenyl nucleotide binding	5.6 x10 <sup>-6</sup>	8.4 x10 <sup>-4</sup>
	BP	0043687	post-translational protein modification	6.2 x10 <sup>-6</sup>	5.3 x10 <sup>-3</sup>
	MF	0001883	purine nucleoside binding	8.2 x10 <sup>-6</sup>	7.4 x10 <sup>-4</sup>
	MF	0001882	nucleoside binding	9.7 x10 <sup>-6</sup>	7.3 x10 <sup>-4</sup>
	MF	0005524	ATP binding	1.5 x10 <sup>-5</sup>	9.6 x10 <sup>-4</sup>
	MF	0032559	adenyl ribonucleotide binding	2.1 x10 <sup>-5</sup>	1.2 x10 <sup>-3</sup>
	MF	0003824	catalytic activity	7.5 x10 <sup>-5</sup>	3.1 x10 <sup>-3</sup>
	BP	0006468	protein amino acid phosphorylation	8.6 x10 <sup>-5</sup>	3.6 x10 <sup>-2</sup>
	BP	0043412	biopolymer modification	1.1 x10 <sup>-4</sup>	3.8 x10 <sup>-2</sup>
	BP	0019538	protein metabolic process	1.4 x10 <sup>-4</sup>	3.4 x10 <sup>-2</sup>
	BP	0006464	protein modification process	2.0 x10 <sup>-4</sup>	3.7 x10 <sup>-2</sup>
	MF	0017076	purine nucleotide binding	2.7 x10 <sup>-4</sup>	8.2 x10 <sup>-3</sup>
	MF	0004672	protein kinase activity	5.9 x10 <sup>-4</sup>	1.4 x10 <sup>-2</sup>
	MF	0032553	ribonucleotide binding	8.0 x10 <sup>-4</sup>	1.7 x10 <sup>-2</sup>
	MF	0032555	purine ribonucleotide binding	8.0 x10 <sup>-4</sup>	1.7 x10 <sup>-2</sup>
	MF	0004713	protein tyrosine kinase activity	1.9 x10 <sup>-3</sup>	3.5 x10 <sup>-2</sup>
	MF	0016301	kinase activity	2.1 x10 <sup>-3</sup>	3.7 x10 <sup>-2</sup>
	MF	0000166	nucleotide binding	2.2 x10 <sup>-3</sup>	3.6 x10 <sup>-2</sup>
MF	0016772	transferase activity, transferring phosphorus-containing groups	2.8 x10 <sup>-3</sup>	4.0 x10 <sup>-2</sup>	
Annotation Cluster 3	MF	0008270	zinc ion binding	7.3 x10 <sup>-4</sup>	1.6 x10 <sup>-2</sup>
	MF	0043169	cation binding	1.9 x10 <sup>-3</sup>	3.6 x10 <sup>-2</sup>
	MF	0046872	metal ion binding	2.3 x10 <sup>-3</sup>	3.6 x10 <sup>-2</sup>
	MF	0043167	ion binding	2.8 x10 <sup>-3</sup>	4.2 x10 <sup>-2</sup>
	MF	0046914	transition metal ion binding	2.9 x10 <sup>-3</sup>	4.0 x10 <sup>-2</sup>

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

## 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., Chereil, 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.
- Wilmington, 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.

## Chapter 4

# Protein products of tissue-specific alternative splicing

### 4.1 Introduction

In the previous chapter, I have described evolutionary mechanisms that can increase diversity in the proteome of an organism. Cellular processes that I have addressed there can modulate a protein's role in a cell by adding novel functional segments to the ancestral proteins. I have also discussed there the potentially important role of intergenic alternative splicing in protein evolution. Intergenic splicing can be an intermediate step in gene fusion, and after gene fusion, alternative splicing can enable expression of both the ancestral protein variant, and a novel protein with a gained protein domain. Moreover, because of alternative splicing, many genes in the higher eukaryotic genomes are able to express a number of different protein products. Thus, for example, there are on average four isoforms for every gene in the human genome (Jin et al., 2004). Protein isoforms produced by alternative splicing increase protein diversity. Additionally, a particular isoform can modulate processes different to those modulated by other products of the same gene. Expression of these isoforms, that have a function distinct from other products of the same gene, is likely to be carefully regulated.

It is well known that the same gene can be used in more than one signalling pathway. Sometimes, for example in the case of genes involved in the well studied extracellular signal-regulated kinase (ERK) cascade of the mitogen-activated protein kinase (MAPK) pathway, regulated cellular processes can be as distinct as proliferation, differentiation, apoptosis, learning and memory (Shaul et al., 2009). Nonetheless, central genes in this cascade, such as MEK and ERK, play a crucial role independently of the process that will eventually be induced. The position of these genes in the ERK cascade is illustrated in Figure 4.1. Thus, one of the fundamental questions is how fidelity in signalling is achieved, as it is clear that other regulatory mechanisms, apart from the sole level of gene expression, are necessary for attainment of the specific cellular response. One level of regulation is expression of different protein isoforms (Shaul and Seger, 2007). For example, in the MAPK pathway, the interaction of specific alternative splice forms of the ERK1 and MEK1 genes facilitates mitotic Golgi fragmentation while interaction of other ERK1 and MEK1 splice forms plays a role in the response to growth factor signals (Shaul and Seger, 2007).

Here, I investigate the hypothesis that, due to alternative splicing, genes that are used in different cellular networks often express protein isoforms with distinct binding motifs. Exposition of different binding peptides would provide a powerful mechanism for enabling the same gene to function in different cellular pathways. Moreover, it is likely that these differentially expressed binding peptides lie in disordered protein regions. There are several reasons for proposing this. Firstly, disordered protein regions are known to play crucial roles in regulation and signalling (Gspöner and Babu, 2009; Gspöner et al., 2008; Wright and Dyson, 1999). Furthermore, these regions are preferred over structured protein segments in protein-protein interactions (PPI) (Shimizu and Toh, 2009) and are abundant in hub proteins of higher eukaryotes (Dosztanyi et al., 2006; Haynes et al., 2006). Finally, alternative inclusion of short disordered regions is less likely to disrupt the overall protein structure (Romero et al., 2006).

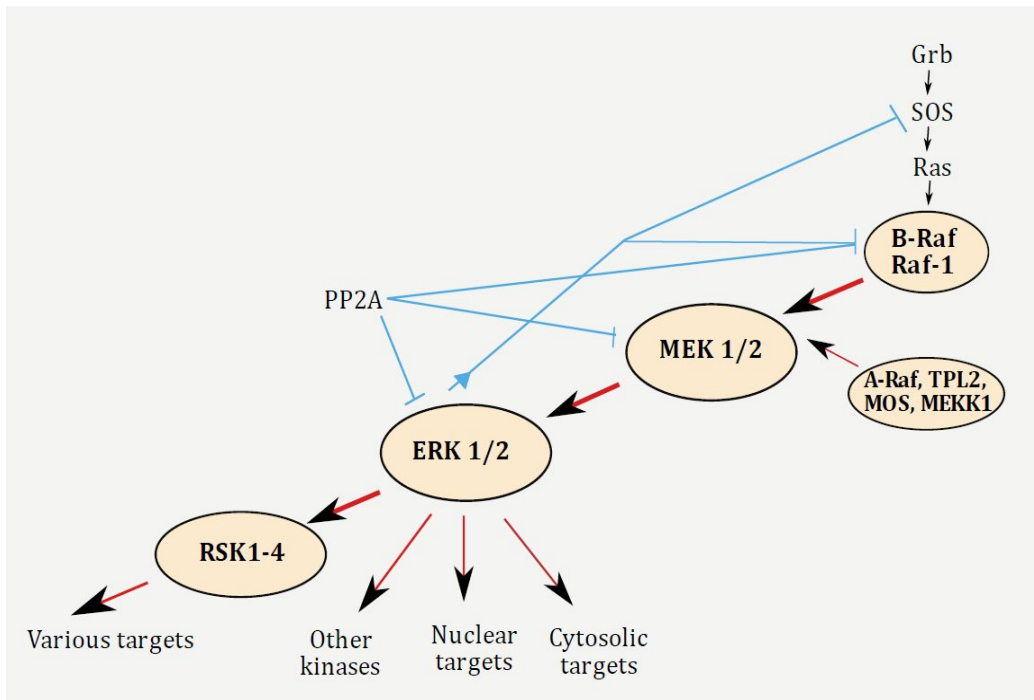


Figure 4.1: Schematic representation of the ERK signaling cascade. The bold arrows show the main pathway upon growth factor activation. Red arrows show activatory phosphorylation events, green accessory phosphorylation and blue inhibitory phosphorylation and dephosphorylation. The illustration is adapted from (Shaul et al., 2007).

Previous studies of alternative splicing at the protein level have shown that the residues that are differentially present between the splice isoforms frequently fall in the intrinsically disordered protein regions (Romero et al., 2006). This can be a consequence of avoidance of structured protein domains, but could also imply a connection between the individual isoform and specific function. If alternative inclusion of protein segments with distinct binding motifs is used to modify the behaviour of the protein in cellular pathways, then this process should be carefully regulated. Hence, protein segments encoded by finely regulated alternative splicing are more likely to be enriched in functionally significant regions compared to all other alternatively spliced segments. The structure of a protein with both ordered and disordered regions is illustrated in Figure 4.2.



Figure 4.2: The structure of a human mitochondrial protein apoCox17 illustrates a protein with both ordered and disordered regions. The change of the structure colour from blue to red indicates direction of the sequence from N- to C-terminus. The positions of amino acids at the disordered N-terminus (blue) are flexible and cannot be clearly defined in the structure. The illustration is taken from the PDB database ([www.pdb.org](http://www.pdb.org)).

Wang et al. recently reported a set of human exons that were differentially expressed between different tissues (Wang et al., 2008). In their study, Illumina deep sequencing of complementary DNA fragments was used to assess the level of alternative splicing in the human genome. Ten different human tissues and five different cell lines were used in the study: adipose, brain, breast, cerebellum, colon, heart, liver, lymph node, skeletal muscle, and testes; BT474, HME, MB435, MCF7 and T47D. Tissue-specific expression was assessed by comparing read data in each tissue sample to that in the other. Since tissue-samples were taken from different individuals, a portion of the differentially expressed exons might have represented allele specific splicing. The authors addressed this issue by comparing samples from the same tissue – cerebellar cortex – between different individuals and showed that the main difference in exon expression was indeed due to tissue-specific splicing regulation.

In this chapter, I discuss the function of tissue-specifically expressed protein segments and the possible role that these regions have in regulation of processes in the tissues where they are expressed. I investigate a hypothesis that in humans, and most likely higher eukaryotes in general, protein functions in different tissues can expand through alternative inclusion of functional disordered segments. In this way, the same gene could be used in different cellular pathways.

## 4.2 Methods

### 4.2.1 Sets of tissue-specific, cassette and constitutive exons

Co-ordinates of tissue-specific exons were obtained from the study by Wang and colleagues (Wang et al., 2008) and then mapped to the longest Ensembl transcripts (Ensembl release 54) where the difference between these coordinates and the coordinates of known Ensembl exons was at most two nucleotides. Next, sets of cassette and constitutive exons were composed for a comparison (Figure 4.3). The set of cassette exons was composed from all cassette Ensembl exons. The aim here was to follow the rationale of the ASTD database (Koscielny et al., 2009) in classifying cassette exons and include in the set those instances where an entire exon was either present or absent in at least two transcripts. Finally, each gene in Ensembl 54 was represented with the longest transcripts it encodes. All other exons in the representative transcripts, which did not overlap with tissue-specific or cassette exons, made a set of constitutive exons. It is important to note that the annotation of an exon as any of these three types does not necessarily describe the exon correctly. For example, exons classified as cassette exons likely contain tissue-specific exons that have not been reported in the study by Wang et al., which is used here as a reference for tissue-specific exons. Furthermore, among the constitutive exons are most likely also the cases of exons that are differentially included in different isoforms, but not all gene isoforms have been experimentally verified yet. Finally, as indicated by the study by Wang et al. a list of all exons in the human genome is still far from being complete. Only the transcripts with two or more exons were considered in the analysis and a script was used to map exon borders to the



corresponding protein coding sequences. Information about exon borders was obtained through the Ensembl BioMart and API.

#### 4.2.2 Enrichment of genes with specific function in the set of tissue-specific exons

The DAVID service (Dennis et al., 2003) was used to investigate whether genes that were reported to have a tissue-specific exon, which also mapped to a known Ensembl exon, were enriched in any molecular function GO terms. Genes with tissue-specific exons were uploaded and compared against the database background of human genes. The DAVID service was also used to test over-representation of specific BioCarta cellular pathways in the set of tissue-specific genes.

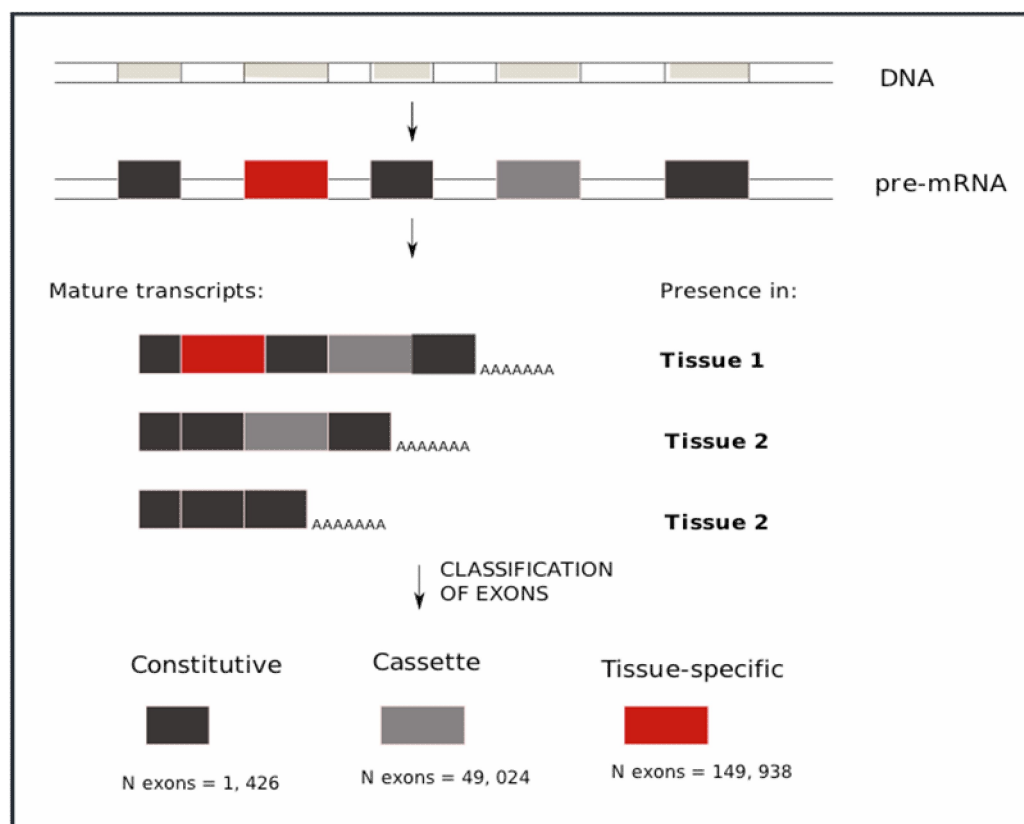


Figure 4.3: Scheme of exon classification. Tissue-specific exons are obtained from the study by Wang et al., while sets of cassette and constitutive exons are made by classifying Ensembl coding exons according to this scheme.

### 4.2.3 Prediction of disordered protein residues

Disordered regions were predicted for protein sequences of the representative transcripts that contained previously described tissue-specific, cassette or constitutive exons, using the IUPred (Dosztanyi et al., 2005) and VSL2B (Peng et al., 2006) software. The IUPred software predicts unstructured protein regions based on the lack of favourable interactions between adjacent amino acids. VSL2B is a baseline predictor of the VSL2 method, which uses a support vector machine method for prediction of disordered residues. VSL2B takes into account only the amino acid composition of a protein, and since it is faster than VSL2 it is recommended for genome-scale studies (Peng et al., 2006). This prediction method recognizes only the symbols for the standard 20 amino acids, so all non-standard symbols (positions with ambiguously assigned amino acids) were removed from the sequences and after the prediction was carried out the removed amino acids were assigned the same status that the surrounding amino acids were predicted to have (disorder or order).

### 4.2.4 Prediction of functional residues

Binding motifs in the sequences of all proteins included in the study were predicted using the ANCHOR software. When it was possible to find the identical protein sequence for the proteins from this study in the Swiss-Prot section of the UniProt database (UniProt release 15.5, which was in concordance with the Ensembl version 54), the Ensembl transcript identifiers from this study were mapped to the corresponding UniProt protein identifiers, and information about the positions of post-translationally modified (PTM) sites in these proteins was obtained. PTM sites that were included in the analysis were: phosphorylation, methylation, acetylation, amidation, addition of pyrrolidone carboxylic acid, isomerisation, hydroxylation, sulfation, flavin-binding, cystein oxidation and nitrosylation sites. When it was possible to find the corresponding international protein index - IPI identifier - for proteins in this study in the Phosida database,

positions of experimentally predicted phosphosites were mapped onto proteins. It was required that proteins analysed in this study contained the reported Phosida phosphopeptides.

#### 4.2.5 Conservation of exons in the three different datasets

The representative genes with exons from the tissue-specific, cassette or constitutive set were mapped to orthologous mouse genes using the Galaxy service (Taylor et al., 2007). It was investigated whether the mouse genome had regions homologous to the exons from this study, and when the homologous regions were present, the level of similarity between them was assessed. Mouse sequences that are orthologous to the exons in these three sets were downloaded from the Galaxy website - pairwise alignments for human genome 18 and *Mus musculus* 9 were used in the study. Fractions of identical aligned nucleotides per exon in the three sets were calculated. The same analysis was performed for aligned disordered residues only - those predicted by IUPred - and for aligned binding peptides only - those predicted by ANCHOR. Additionally, for each set of exons, a fraction of all coding residues for which it was possible to extract the orthologous mouse sequence was calculated. Similarly, a fraction of disordered residues and of the residues in the binding peptides for which it was possible to extract the orthologous sequence was calculated.

#### 4.2.6 Significance of observed trends

To test whether the differences in the fractions of disordered residues, predicted binding motifs, annotated PTM sites and experimentally predicted phosphosites in the three sets of exons were significant Chi-square tests were applied by using the R software. Significance of exon and peptide conservation in the tissue-specific set compared to two other sets, as well as conservation of peptide versus all other residues in the tissue-specific set, were tested with the Mann-Whitney test (Wilcox test in the R software). The Mann-Whitney test was applied because

the distribution of exon conservation values did not follow the normal distribution ( $P < 2.2 \times 10^{-16}$ , Shapiro-Wilk test for the distribution of values for tissue-specific exons). Test sets of cassette and constitutive exons with the same average length as in the set of tissue-specific exons were composed and fractions of predicted binding motifs and annotated PTM sites were calculated. The significance in the difference of fractions of the predicted functional residues was tested with the Chi-square test, again using the R software.

#### 4.2.7 Comparison of MEK1 and MEK2 protein sequences

Mouse MEK1 and MEK2 protein sequences were downloaded from the Ensembl database. Proteins were aligned using the Needleman-Wunsch algorithm (with a gap opening cost of 10.0 a and gap extension cost of 0.5) from the EBI online service ([www.ebi.ac.uk/Tools/emboss/align/index.html](http://www.ebi.ac.uk/Tools/emboss/align/index.html)). Disordered residues were predicted in these sequences with the IUPred software and fractions of disordered residues between the aligned and unaligned protein segments were calculated.

#### 4.2.8 Enrichment of known disease genes in the set of tissue-specific exons

Genes with phenotype annotations and assigned human homologues were downloaded from the Mouse Genome Informatics database. The significance in the fraction of genes with tissue-specific isoforms among the genes related to embryonic lethality was tested with the ChiSquare test, using the R software. Cancer gene census (downloaded on 21 Sep 2009) and genes from the COSMIC database (release 43) were downloaded from the corresponding databases. Genes with tissue-specific variants and the background set of all human genes in the Ensembl version 54 were mapped to their human gene nomenclature identifiers, using the Ensembl API. The significance in the fraction of disease causing genes between the two sets of genes was calculated again with the Chi-square test.

## 4.2.9 Disorder signatures in the protein products of the p73 gene

The protein sequence of the longest isoform of the p73 gene, TP73-001, was taken from the Ensembl database. Disorder and binding peptides were predicted using the IUPred and ANCHOR online services, respectively.

## 4.3 Results

### 4.3.1 Sets of exons with different expression profiles

I investigated whether genes with protein coding tissue-specific exons were associated with any particular molecular function. I found that these genes were enriched with protein-binding, transferase and kinase activity GO terms (Table 4.1). Hence, it is possible that they mediate processes which in different tissues include different protein partners. One possibility for achieving this is through utilization of functional disordered protein segments.

To test this hypothesis, I analysed three different sets of exons: (i) Protein coding exons that map to known Ensembl (Hubbard et al., 2009) transcripts and are differentially expressed between at least two different tissues or cell lines (tissue-specific exons), as reported by Wang et al. (Wang et al., 2008). (ii) Coding exons that differ in whether they are present or absent between at least two transcripts of the same gene (cassette exons), as annotated in Ensembl. I excluded from this set those exons that overlapped with other cassette exons or with the tissue-specific exons. (iii) Coding exons that cannot be classified as alternatively spliced according to the current Ensembl gene annotations (constitutive exons). There were 1 426 tissue-specific, 49 024 cassette and 149 938 constitutive coding exons in their respective sets. Figure 4.3 illustrates the classification scheme.

Table 4.1: Significant molecular function GO terms enriched in the genes with tissue-specific exons (P-value < 0.05). Subset of significantly enriched molecular function GO terms in the set of genes with tissue-specific exons (P-value < 0.1). EASE P-values represent modified Fisher exact P-values (Hosack et al., 2003). Column 'Benjamini' shows P-values after applying the Benjamini correction for multiple tests.

GO term description	GO term ID	EASE P-value	Benjamini P-value
Protein binding	0005515	$5.4 \times 10^{-16}$	$1.6 \times 10^{-12}$
Cytoskeletal protein binding	0008092	$9.7 \times 10^{-13}$	$1.4 \times 10^{-9}$
Actin binding	0003779	$4.3 \times 10^{-10}$	$4.1 \times 10^{-7}$
Binding	0005488	$1.1 \times 10^{-5}$	$7.7 \times 10^{-3}$
Catalytic activity	0003824	$1.8 \times 10^{-5}$	$1.0 \times 10^{-2}$
Transferase activity	0016740	$2.1 \times 10^{-5}$	$1.0 \times 10^{-2}$
Transferase activity, transferring phosphorus-containing groups	0016772	$3.2 \times 10^{-5}$	$1.3 \times 10^{-2}$
Kinase activity	0016301	$4.4 \times 10^{-5}$	$1.6 \times 10^{-2}$
Protein serine/threonine kinase activity	0004674	$5.3 \times 10^{-5}$	$1.7 \times 10^{-2}$
Enzyme binding	0019899	$1.1 \times 10^{-4}$	$3.2 \times 10^{-2}$
Nucleotide binding	0000166	$1.3 \times 10^{-4}$	$3.4 \times 10^{-2}$
Ras GTPase binding	0017016	$1.9 \times 10^{-4}$	$4.3 \times 10^{-2}$

### 4.3.2 Tissue-specific exons are enriched in disordered residues

I compared the fractions of disordered residues in the three sets of exons with different expression profiles. Figure 4.2 shows a protein which contains both ordered and disordered regions. Disordered regions were identified using the IUPred software (Dosztanyi et al., 2005), which predicts unstructured protein regions in the segments with biased amino acid composition, such as those enriched in polar or charged residues, which do not allow formation of sufficient stabilizing interactions. I found that both sets of alternatively spliced exons - the set of tissue-specific and the set of cassette exons - were enriched with exons encoding disordered amino acids, when compared to the set of constitutive exons (Figure 4.4). The fraction of exons coding for unstructured protein regions was the highest for the tissue-specific exons (31% of tissue-specific exons were predicted to have 50% or more disordered residues, compared to 25 and 16% of cassette and constitutive exons, respectively). The difference in the number of disordered exons was significant when tissue-specific exons were compared to both cassette and constitutive exons ( $P < 5.1 \times 10^{-7}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Chi-square test, where the value of  $2.2 \times 10^{-16}$  is the smallest P-value in R). Furthermore, to investigate whether protein disorder is in general a feature of genes that undergo tissue-specific splicing or if it is a specific characteristic of tissue-specific exons, I compared the fraction of disordered residues among the tissue-specific exons to the fraction of disordered residues in all other exons encoded by the representative transcripts with these exons. This showed that disordered residues are indeed characteristic for alternatively spliced tissue-specific exons (444 out of 1426 tissue-specific exons were encoding mostly disordered protein segment, compared to 3,543 out of 16,850 all other exons in these transcripts,  $P < 2.2 \times 10^{-16}$ , Chi-square test).

To ensure that observations about the prevalence of disordered residues in the tissue-specific exons are not biased by the applied disorder prediction method I used another method for identification of disordered regions. The VSL2B software is trained on datasets of disordered proteins and uses a linear support vector machine approach based on amino acid composition. Prediction

of intrinsically disordered residues by this method confirmed that disordered residues are most common in the set of tissue-specific exons, followed by cassette exons. The fractions of exons with at least 50% predicted disordered residues were 53, 46 and 36% in the sets of tissue-specific, cassette and constitutive exons, respectively (Table 4.2). Hence, the observed enrichment of tissue-specific exons in disordered regions seems to be independent of the method for disorder prediction.

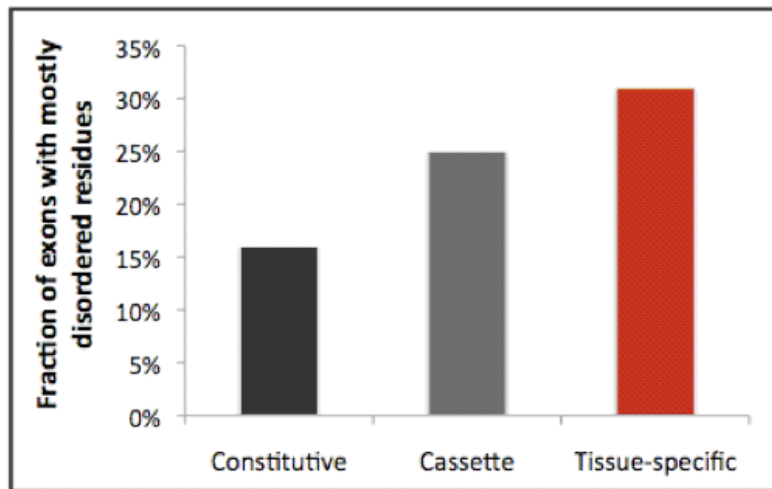


Figure 4.4: Protein regions encoded by tissue-specific exons are enriched in intrinsically disordered residues. The fraction of exons with at least 50% disordered residues in the three different sets of exons is shown. The number of exons with mostly disordered residues was significantly higher among tissue-specific exons when compared to cassette and constitutive exons ( $P=5.1 \times 10^{-7}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Chi-square test, details in Appendix C.1). Disordered residues were predicted with the IUPred software.



Table 4.2: Fractions of exons with at least 50% disordered residues, as predicted by the VSL2B software. The fraction of exons with mostly disordered residues is still predicted to be the highest in the set of tissue-specific exons followed by the cassette exons. The column P-value shows significance of this enrichment compared to the two other sets of exons as calculated by the Chi-square test.

Analysis	Set of exons	Fraction of disordered exons	P-value
VSL2	Tissue-specific	53%	/
	Cassette	46%	$P < 5.4 \times 10^{-8}$
	Constitutive	36%	$P < 2.2 \times 10^{-16}$

### 4.3.3 Functional residues in disordered segments encoded by tissue-specific exons

My hypothesis in this study is that disordered regions encoded by tissue-specific exons expose functional protein segments (Romero et al., 2006). Alternatively, these regions could act as fillers between functional structured domains (Tress et al., 2008; Tress et al., 2007). Functional disordered residues are frequently used in transient interactions in the cell, since their intrinsic flexibility allows them to be readily accessible to the proteins they interact with (Gsponer and Babu, 2009). I investigated here whether tissue-specific disordered residues indeed encode segments that could be used in protein interactions. Possible short protein binding sites and sites of post-translational modifications (PTMs) reflect disordered protein regions. Here, I analyzed whether there is evidence for a connection between tissue-specific disordered regions and protein binding sites. Firstly, I investigated whether unstructured segments contained peptide motifs that were likely to be bound by other proteins. For this, I used the ANCHOR software (Meszaros et al., 2009), which identifies disordered regions with a potential to bind protein domains on the hypothetical interaction partners. I found enrichment for the predicted functional peptide motifs in the tissue-specific exons compared to cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$

and  $P < 2.2 \times 10^{-16}$ , respectively, Chi-square test). Among the tissue-specific exons, 29% had a binding motif, compared to 18% of cassette exons and 18% of constitutive exons, see Figure 4.5a.

In addition, I investigated whether PTM sites were enriched in tissue-specific exons. For this, I looked at the annotated PTM sites in the Swiss-Prot portion of the UniProt database (Consortium, 2009). The analysis covered phosphorylation, methylation, acetylation and other PTM sites (Methods). This revealed that enrichment of PTM sites was indeed present in the set of tissue-specific exons. Tissue-specific exons had significantly more predicted PTM sites than cassette and constitutive exons ( $P < 9.9 \times 10^{-12}$  and  $P < 3.2 \times 10^{-7}$ , respectively, Chi-square test). Among the tissue-specific exons from those transcripts that were successfully mapped to the UniProt isoforms, 13% had a PTM, compared to 7 and 8% of cassette and constitutive exons, respectively, see Figure 4.5b. PTM sites are frequently associated with unstructured regions (Holt et al., 2009; Iakoucheva et al., 2004) and in the set of tissue-specific exons, the majority (69%) of exons with at least one PTM site had a PTM in the predicted disordered region.

As a control, I investigated if the same signal could be detected for an independent set of experimentally identified PTM sites. For this, I used the information about human phosphorylation sites stored in the Phosida database (Gnad et al., 2007). These data came from the mass spectrometry experiment that studied phosphorylation sites in HeLa cells in their basal state and upon stimulation with the epidermal growth factor (Olsen et al., 2006). I computed the fraction of exons with Phosida phosphosite(s) in each of the three sets of exons and found that tissue-specific exons had a significantly higher fraction of phosphosites compared to cassette and constitutive exons (Table 4.3). Taken together, several independent analyses confirmed that the set of tissue-specific exons is enriched in functionally annotated sites associated with disorder.

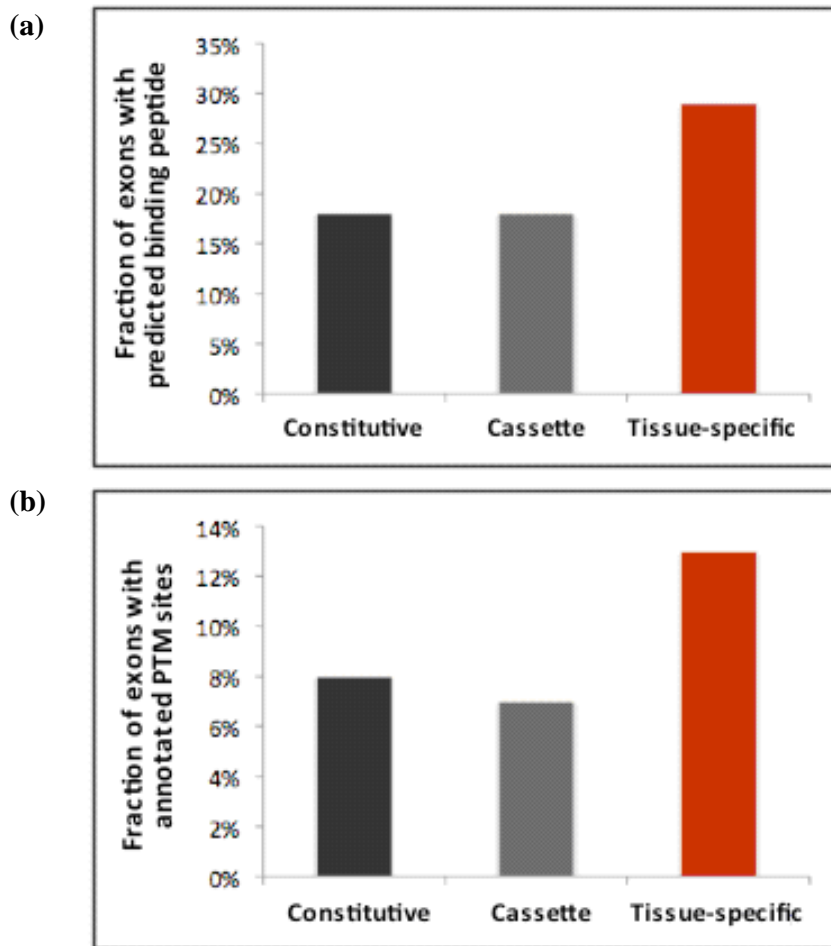


Figure 4.5: Tissue-specific exons encode protein segments enriched with predicted binding motifs and annotated PTM sites. (a) Fraction of exons with encoded binding motifs in the three different sets of exons. Binding motifs were predicted with the ANCHOR software. Tissue-specific exons were found to have a significantly higher fraction of predicted binding motifs than cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Chi-square test, details in Appendix C.1). (b) Fraction of exons with annotated PTM sites in the three different sets of exons. Tissue-specific exons were found to have a significantly higher fraction of PTM sites than cassette and constitutive exons ( $P \leq 9.9 \times 10^{-12}$  and  $P \leq 3.2 \times 10^{-7}$ , respectively, Chi-square test, details in Appendix C.1). Positions of PTM sites in proteins were taken from the Swiss-Prot portion of the UniProt database.

Table 4.3: Fractions of exons with a phosphosite identified in a single large scale experiment (Olsen et al., 2006). The fraction of exons with a phosphosite is the highest in the set of tissue-specific exons followed by the constitutive exons. The column headed P-value shows the significance of the enrichment in tissue-specific exons compared to the two other sets as calculated by a Chi-square test.

Analysis	Set of exons	Fraction of exons with a phosphosite	P-value
Phosida phosphosites	Tissue-specific	2.3%	/
	Cassette	0.4%	$P < 2.2 \times 10^{-16}$
	Constitutive	0.5%	$P < 2.2 \times 10^{-16}$

#### 4.3.4 Distribution of functional residues in the control sets of cassette and constitutive exons

Comparison of average exon lengths in the three sets of exons showed that tissue-specific exons were on average longer than cassette and constitutive exons; the average length of tissue-specific exons was 68 nucleotides (close to 23 amino acids), and the average lengths of cassette and constitutive exons were 46 and 54 nucleotides (15 and 18 amino acids, respectively). The fraction of exons with a predicted binding peptide or PTM site could be influenced by the length of tested exons. Therefore, I investigated if the difference in exon lengths affected the results which indicated enrichment for functional sites in the tissue-specific exons.

I filtered out shorter exons from the sets of cassette and constitutive exons in order to compose test sets with the average length of exons of 68 nucleotides. I compared the fractions of exons with predicted binding peptides and PTM sites in these two test sets with the one in the set of tissue-specific exons. I found that the set of tissue-specific exons still encoded a significantly higher fraction of PTM sites than the two test sets (Table 4.4). With regard to predicted binding peptides, I found that their fraction was significantly higher

among the tissue-specific exons when compared to constitutive exons, but the difference was not that dramatic when compared to cassette exons (Table 4.4). Cassette exons have a higher fraction of disordered regions, so in that sense, it is not surprising that disordered binding motifs are more frequently predicted in that set than in the set of constitutive exons. However, overall, the analysis of subsets with longer cassette and constitutive exons confirmed that the observed enrichment of tissue-specific exons with functional sites is independent of the exon length.

Table 4.4: Fractions of exons with either a predicted binding peptide or an annotated PTM site in the sets of tissue-specific exons and in the sets of cassette and constitutive exons that are filtered to have the same average length as tissue-specific exons. The column P-value shows the significance of the enrichment of tissue-specific exons with these functional sites compared to the two other sets as calculated by Chi-square test.

Analysis	Set of exons	Fraction of exons with a functional site	P-value
Binding peptides	Tissue-specific	29%	N/A
	Cassette	26%	$P=2.9 \times 10^{-2}$
	Constitutive	23%	$P=2.0 \times 10^{-8}$
PTM sites	Tissue-specific	13%	N/A
	Cassette	8%	$P=1.6 \times 10^{-8}$
	Constitutive	8%	$P=3.3 \times 10^{-7}$

### 4.3.5 Disordered residues encoded by tissue-specific exons are highly conserved

While the tissue-specific unstructured protein regions show apparently enrichment for binding motifs and PTM sites, it is known that unstructured proteins generally evolve faster than the structured ones (Brown et al., 2002). Hence, such peptide motifs could have occurred by chance. However, if they are functionally relevant then it is more likely that the unstructured regions and the predicted peptide motifs will be evolutionary conserved. Therefore, I investigated the similarity of exons from the three different sets with orthologous sequences in mouse. I compared the fractions of identical aligned nucleotides per exon in the three sets of exons and found that tissue-specific exons were significantly more conserved than cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Mann-Whitney test, Table 4.5).

I performed the same analysis for aligned disordered regions in the exons only. Again, I found that residues in disordered regions in tissue-specific exons were significantly more conserved than those in disordered regions of cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Mann-Whitney test). The difference in the conservation of disordered regions was even more dramatic than the difference in the conservation of all residues in these three sets of exons (Table 4.5). The median value of conservation for residues in disordered segments was 0.90 in tissue-specific exons, 0.83 in cassette exons and 0.84 in constitutive exons (Figure 4.6).

Next, I looked at the conservation of predicted binding peptides only. Conservation of binding peptides was higher than the overall conservation of exons in all three sets, and it was the highest in the set of tissue-specific exons. Importantly, predicted binding residues were not only significantly more conserved in the tissue-specific exons when compared to cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Mann-Whitney test, Table 4.5), but were significantly more conserved than all other residues in the tissue-specific exons alone ( $P = 6.3 \times 10^{-6}$ , Mann-Whitney test, Table 4.6). Thus, even though the binding function of these residues is only predicted, it is likely that they play an important role in these proteins. The median value of

conserved predicted binding peptides was 0.91 in tissue-specific exons, 0.86 in cassette exons and 0.86 in constitutive exons (Figure 4.6).

For some residues, or whole exons, it was not possible to extract the orthologous mouse sequence and the reason for this is either that there is no orthologous sequence in mouse or that the two regions have evolved beyond recognition. If I take into account information about residues for which it was possible to extract the orthologous sequence, the observed high conservation of tissue-specific exons becomes even more prominent. Namely, I was able to extract the orthologous sequence for 98% of residues in tissue-specific exons, 91% in cassette and 96% of residues in constitutive exons. Since disordered residues evolve in general faster, it is not surprising that on average less of them had a corresponding orthologous sequence: 98% of disordered residues in tissue-specific exons, 87% in cassette and 94% of residues in constitutive exons were aligned with their mouse orthologous sequence. Hence, this observation also confirms high conservation of the whole exons and in particular of the residues encoding disordered segments in the set of tissue-specific exons.

Taken together, the observed evolutionary conservation of tissue-specific exons likely reflects a functional constraint, which could have emerged due to functionally important peptide motifs.

Table 4.5: Conservation of exons in different sets, and of different elements in these exons. The number of exons encoding disordered segments and binding peptides for which orthologous mouse sequences were found is indicated in the column  $N_{\text{exons}}$ . The column headed Median shows the median value for the fractions of nucleotides in each exon that are identical to the aligned mouse nucleotides. The column P-value shows the significance of the difference in conservation between the set of tissue-specific exons and each of the two other sets as calculated by the Mann-Whitney test.

Set for analysis	Set of exons	$N_{\text{exons}}$	Median	P-value
Whole exons	Tissue-specific	1,404	0.89	N/A
	Cassette	44,750	0.86	$P < 2.2 \times 10^{-16}$
	Constitutive	143,811	0.87	$P < 2.2 \times 10^{-16}$
Disordered regions	Tissue-specific	883	0.90	N/A
	Cassette	24,120	0.83	$P < 2.2 \times 10^{-16}$
	Constitutive	68,719	0.84	$P < 2.2 \times 10^{-16}$
Binding peptides	Tissue-specific	630	0.91	N/A
	Cassette	13,600	0.86	$P < 2.2 \times 10^{-16}$
	Constitutive	37,708	0.86	$P < 2.2 \times 10^{-16}$

Table 4.6: Predicted binding peptide sites in Tissue-specific exons are significantly more conserved than other residues in these exons. The P-value is calculated with the Mann-Whitney test. The number of exons that were applicable for the test is shown in the column  $N_{\text{exons}}$ . The column Median shows the median for conservation of binding peptide residues or all other residues in the tissue-specific exons.

Set for analysis	Residues	$N_{\text{exons}}$	Median	P-value
Tissue-specific exons	Binding peptides	630	0.91	$P < 26.3 \times 10^{-6}$
	Other	1,363	0.89	



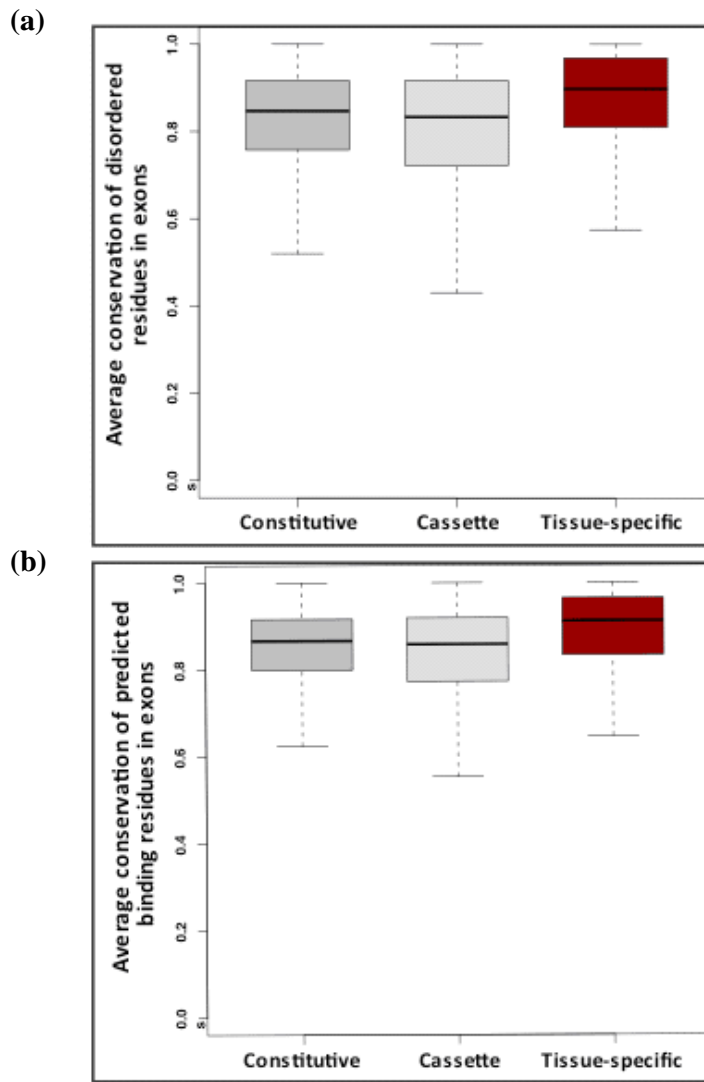


Figure 4.6: Residues in predicted disordered regions and peptide binding sites in the tissue-specific exons are highly conserved. (a) Conservation of predicted disordered residues in the three sets of exons, as calculated from residues aligned with mouse orthologous sequences. The median value for each set is shown as thick black line. Boxes enclose values between the first and third quartile. The interquartile range (IQR) is calculated by subtracting the first quartile from the third quartile and all values that lie more than 1.5x IQR lower than the first quartile or 1.5x higher than the third quartile are considered to be outliers and are not shown on these graphs. The smallest and the highest value that is not an outlier are connected with the dashed line. Disordered residues in tissue-specific exons were found to be significantly more conserved than those in cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Mann-Whitney test, details in Table S3) (b) Conservation of predicted binding peptides, as calculated from residues aligned with mouse orthologous sequences. Predicted binding peptides in tissue-specific exons were found to be significantly more conserved than those in cassette and constitutive exons ( $P < 2.2 \times 10^{-16}$  and  $P < 2.2 \times 10^{-16}$ , respectively, Mann-Whitney test, details in Table S3).

### 4.3.6 Genes with tissue-specifically regulated exons have an important function in organism development and survival

If genes with tissue-specific isoforms tend to take part in different cellular pathways then mutations in these proteins are likely to have severe effects on the cellular and organism phenotype. I performed several analyses to see if this was the case. Firstly, I investigated whether genes from the MGI database (Bult et al., 2008), which are known to cause embryonic lethality in mice when mutated, were enriched with orthologues of human genes that have tissue-specific isoforms. I indeed found that genes with the tissue-specific isoforms were overrepresented among the genes involved in embryonic lethality ( $P < 1.2 \times 10^{-8}$ , Chi-square test, Table 4.7, Figure 4.7), which implied their potentially important role in the early stages of development.

Secondly, I investigated whether mutations in these genes could be related to cancer phenotype, since disruption of signalling pathways is a common initiator of the disease. Moreover, the study by Wang et al. that reported tissue-specific exons also included five different cancer cell lines, which increased the chances of detecting genes whose isoforms were potentially related to cancer. Indeed, I found that both Cancer Gene Census genes (Futreal et al., 2004) (genes that have been causally implicated in cancer) and genes from the COSMIC database (Forbes et al., 2008) (genes found to be somatically mutated in different cancer cells) were enriched with genes that have tissue-specific isoforms (P-values were  $6.2 \times 10^{-2}$  and  $3.2 \times 10^{-6}$  respectively, Chi-square test, Table 4.7, Figure 4.7). This suggested a possible connection between the genes with tissue-specific isoforms and cancer phenotype.

Finally, I investigated whether the genes with tissue-specific isoforms were enriched in any particular cellular pathway since this could possibly imply their influence on the phenotype. I found that these genes were enriched with genes that belong to the PDZ pathway (Table 4.8), a pathway in which disordered residues are known to play an important role. Apart from the significant overrepresentation of genes from PDZ pathway, this analysis revealed another important link; clustering of genes with similar function showed overrepresentation of genes from the MAPK pathway (Table 4.8). A

possible connection with disordered residues here is suggested by the following example from the literature. The MAPK kinase MEK exists in two gene copies, MEK1 and MEK2, which have essentially identical sequences but significantly different effects on the phenotype. I looked at the predicted disordered residues in these proteins and found that 54% of amino acids that differed between MEK1 and MEK2 were predicted to be disordered, compared to only 1% of the identical residues. Therefore, it is possible that in this known example from the MAPK pathway disorder functions as a mediator of protein interactions in a similar way in which I expect it acts in tissue-specific isoforms analysed here.

Taken together, these results suggest that mutations in genes with tissue-specific isoforms can have dramatic effects on the phenotype of an organism by influencing developmental and other crucial signalling pathways and that there is a possible link with disordered residues in the mechanism of its action.

Table 4.7: Genes that are associated with embryonic lethality and cancer phenotype are enriched in genes with tissue-specific isoforms. The  $N_{total}$  column shows the number of genes that I successfully mapped to identifiers in the underlying disease gene databases. The  $N_+$  column shows the number of tissue-specific or all other genes in the databases that are also implicated in disease and  $N_-$  those that are not annotated as such. Background genes in the case of the Mouse Genome Informatics (MGI) database are all human genes with mouse orthologues that have known phenotype effects. In the case of Consensus cancer genes and COSMIC genes, background genes are all human genes in the Ensembl 54 successfully mapped to human gene nomenclature identifiers. Background genes include Tissue-specific genes. P-values are for the Chi-Square tests.

Analysis	Set of genes	$N_+$	$N_-$	$N_{total}$	P-value
MGI	Tissue-specific genes	202	963	1,165	$P < 1.2 \times 10^{-8}$
	All genes in the set	2,080	15,722	17,802	
Consensus cancer genes	Tissue-specific genes	31	1,153	1,184	$P < 6.2 \times 10^{-2}$
	All genes in the set	345	18,630	18,975	
Cosmic	Tissue-specific genes	227	957	1,184	$P < 3.2 \times 10^{-6}$
	All genes in the set	2,697	16,278	18,975	

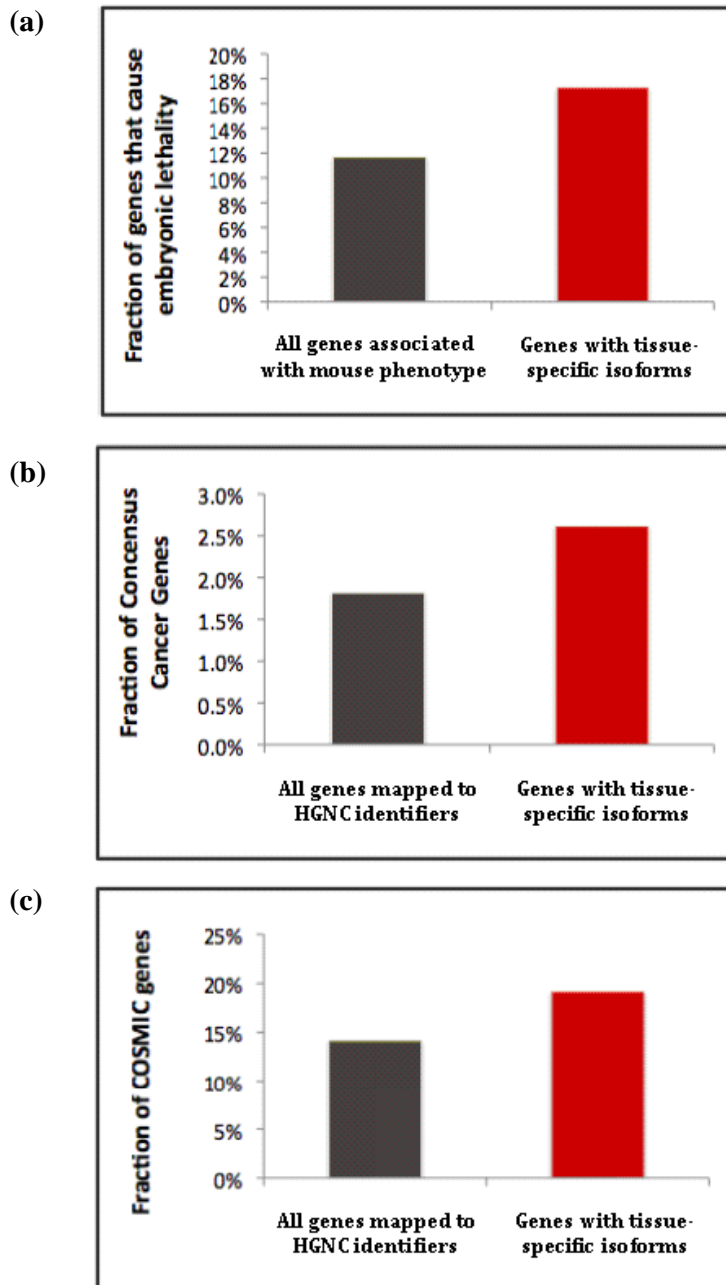


Figure 4.7: Fraction of genes with tissue-specific isoforms that are among the disease causing genes compared to background human genes. This is an illustration of data from Table 4.7. Background genes are always composed of all human genes with the identifiers in the corresponding databases. (a) Fraction of tissue-specific genes (red column) and all genes in the MGI database (grey column) that cause embryonic lethality when mutated. (b) Fraction of tissue-specific genes (red column) and all Ensembl genes with HGNC identifiers that are known to be involved in cancer development. (c) Fraction of tissue-specific genes (red column) and all Ensembl genes with HGNC identifiers that were found to be mutated in cancer but are not necessarily involved in cancer development.

Table 4.8: Pathways overrepresented among the genes with tissue-specific exons. The top results of a search for BIOCARTEA pathways ([www.biocarta.com](http://www.biocarta.com)) that are overrepresented among the genes with tissue-specific exons are shown. Only the most significant individual pathway and cluster of pathways are included in the table. Lists of all terms that are reported to be enriched, but not with high significance are in Appendix C.2. The EASE P-values represent modified Fisher exact P-values (Hosack et al., 2003). The column 'Benjamini' shows P-values after applying the Benjamini correction for multiple tests.

Pathway	EASE P-value	Benjamini P-value
---------	--------------	-------------------

Enriched individual pathway:

PDZ pathway: Synaptic Proteins at the Synaptic Junction	$2.3 \times 10^{-5}$	$7 \times 10^{-3}$
---	----------------------	--------------------

Enriched cluster of pathways with  
similar gene members:

Mapk pathway: MAPKinase Signalling Pathway	0.06	0.90
P38 mapk pathway: p38 MAPK Signalling Pathway	0.26	0.98
Erk Pathway: Erk1/Erk2 MAPK Signalling pathway	0.35	0.99

### 4.3.7 Alternative isoforms of the gene p73

An example from the literature that illustrates the potential importance of alternative inclusion of exons that encode disordered protein segments is the one of the p73 gene. Gene p73 is a homologue of the p53 gene and its main function is tumour suppression. However, this gene encodes a number of splice variants (Figure 4.8) which have been shown to be expressed in a tissue-specific manner (Ishimoto et al., 2002). These different splice isoforms all share the same central DNA binding region and differ in the alternative inclusion of N- and C-terminal exons (Bourdon, 2007). Functionally, the isoforms differ in their binding specificity, and the most striking of them is the  $\Delta Np73$  isoform which lacks the first three exons that encode the 'transactivating region' (Figure 4.8). Instead of acting as a tumour suppressor, the  $\Delta Np73$  isoform acts as an oncogene - possibly by competing with both p53 and other p73 isoforms for the DNA binding site (Ishimoto et al., 2002). When I predicted disordered regions (Dosztanyi et al., 2005) in the main protein isoform TP73001, which includes also the terminal exons, I observed that the encoded protein had several disordered segments and most importantly, that the N-terminal region encoded by the first three exons is predominantly disordered (Figure 4.8). Additionally, this region also contained two predicted binding peptides (not shown), as predicted by ANCHOR (Meszaros et al., 2009). Similarly to the p73 protein, it has been reported previously that the N-terminal region of the human p53 tumour suppressor protein contained large disordered segments (Bell et al., 2002; Dawson et al., 2003). The N-terminal part of the p53 protein has an important regulatory role (Chumakov, 2007), and so far, three different protein partners have been shown to bind to this region - binding peptides for these proteins were successfully predicted with ANCHOR (Meszaros et al., 2009). The example of the p73 gene clearly illustrates that the alternative inclusion of disordered protein segments can dramatically affect the function of a protein.

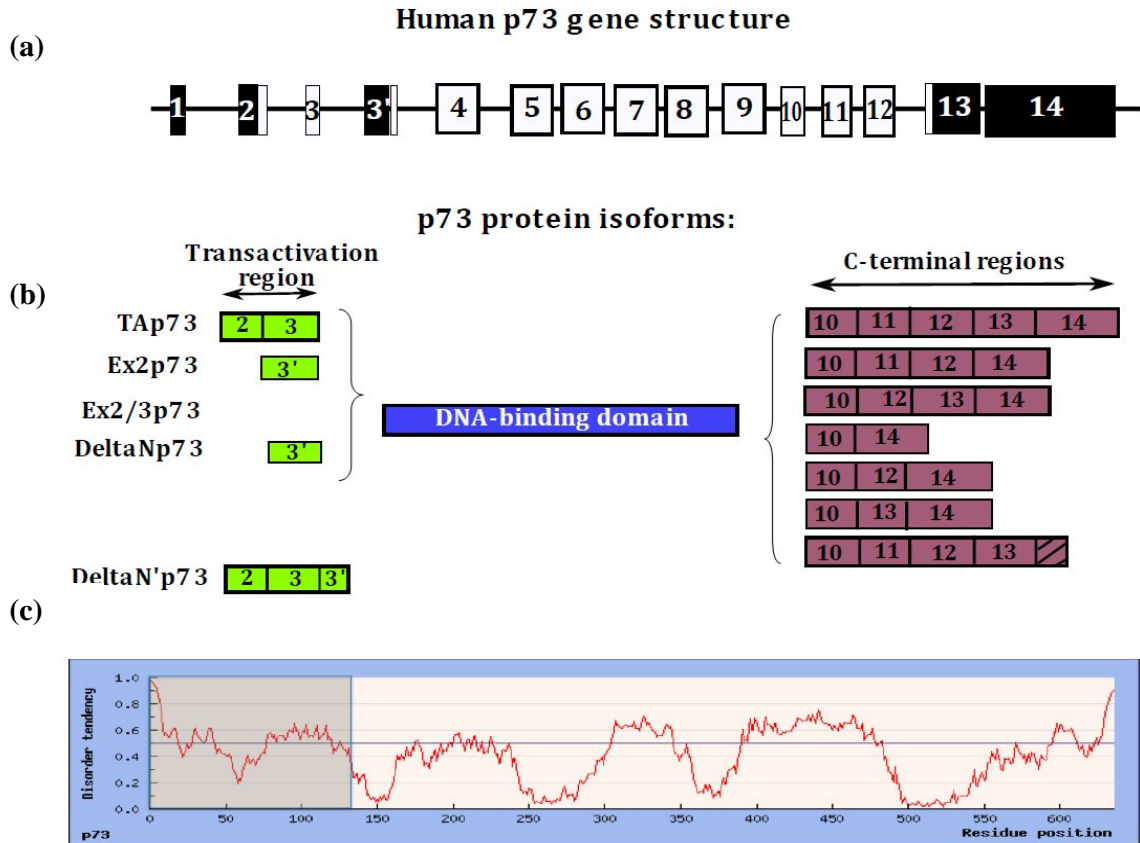


Figure 4.8: p73 gene and its isoforms. Figures a and b are adapted from (Bourdon, 2007), and show different splicing events that occur in the p73 gene. (a) The intron-exon structure of the gene is shown. Black boxes indicate 5' and 3' untranslated exon regions, and white boxes exon regions that encode protein sequences. There are two alternative transcription start sites: before the exon 1 and before the exon 3'. (b) Protein segments encoded by different exons are shown. All splice isoforms apart from  $\Delta N'p73$  have the central DNA-binding domain, but differ in the segments encoded by the N- and C-terminal exons. The exon numbering in section (a) is transferred to section (b) of the figure. (c) Disordered regions (threshold is at 0.5 disorder tendency) in the longest Tap73 (TP73-001) isoform, as predicted with the IUPred software. This isoform includes the first three N-terminal exons that are missing in the  $\Delta Np73$  isoform and these are indicated with a grey square in the disorder prediction graph. The greatest part of the protein segment encoded by these first three exons is predicted to be disordered.



### 4.3.8 Tissue-specific splicing and protein domains

The role of intrinsic disorder as a mediator of protein interactions is becoming increasingly recognized. However, the most studied and better-understood protein interactions are those mediated by protein domains of conserved sequence and defined structure. Therefore, I also investigated whether known protein domains, taken from the Pfam database (Finn et al., 2008), were affected by tissue-specific alternative splicing and if so, what was the predicted function of these domains. It was previously reported that alternative splicing tends to avoid protein domains more frequently than expected by chance (Kriventseva et al., 2003). I investigated if the same trend was present in tissue-specific and cassette exons, and found that indeed domains were avoided in both types of alternative-splicing events. Fractions of exons that overlapped with a predicted Pfam domain (Finn et al., 2008) were 43% and 42% in the sets of tissue-specific and cassette exons, respectively, compared to 54% of constitutive exons that overlapped a Pfam domain (P-value <  $2.47 \times 10^{-15}$  and P-value <  $2.2 \times 10^{-16}$  for tissue-specific and cassette sets of exons, respectively, Chi-square test). This confirmed that alternative splicing tends to avoid protein domains and is more likely to occur in protein regions without annotated domains.

Next, I looked at functional annotation of domains that were completely removed from proteins by tissue-specific alternative splicing. For this, I identified the cases where alternative splicing affected 90% or more of the domain, and exclusion of the tissue-specific exon removed all copies of the domain from a protein. I found that tissue-specific splicing affected predominantly DNA and protein binding domains (Table 4.9). However, this preference for binding domains was not statistically significant. Binding domains are in general common in the human genome, and the similar issue with recognizing the trends that affect these domains has been discussed previously with regard to DNA and protein binding domains in alternative splicing in general (Lareau et al., 2004; Resch et al., 2004). Nonetheless, specific binding domains are likely to play important roles in tissue-specific alternative splicing. An interesting example from Table 4.9 is discussed in Figure 4.9.

Table 4.9: Pfam domains that are removed from the protein products of a gene by tissue-specific alternative splicing. The column Ensembl ID indicates a transcript identifier to which the corresponding tissue-specific exon is mapped, Pfam ID shows the Pfam identifier of the domain that is removed from the protein product and Domain name shows the full name of the affected domain.

General function	Ensembl ID	Pfam ID	Domain name
DNA/RNA binding	ENST00000313565	PF00096	Zinc finger, C2H2 type
	ENST00000235372	PF00096	Zinc finger, C2H2 type
	ENST00000374012	PF00096	Zinc finger, C2H2 type
	ENST00000262965	PF00010	Helix-loop-helix DNA-binding domain
	ENST00000344749	PF00010	Helix-loop-helix DNA-binding domain
	ENST00000378526	PF00645	Polymerase and DNA-Ligase Zn-finger region
	ENST00000380828	PF01754	A20-like zinc finger
	ENST00000321919	PF02178	AT hook motif
	ENST00000257821	PF00628	PHD-finger
	ENST00000389862	PF00035	Double-stranded RNA binding motif
Protein interactions	ENST00000367580	PF07654	Immunoglobulin C1-set domain
	ENST00000400376	PF07686	Immunoglobulin V-set domain
	ENST00000374737	PF00047	Immunoglobulin domain
	ENST00000360141	PF07686	Immunoglobulin V-set domain
	ENST00000356709	PF07686	Immunoglobulin V-set domain
	ENST00000397753	PF00651	BTB/POZ domain
	ENST00000396852	PF02023	SCAN domain
	ENST00000330501	PF02023	SCAN domain
	ENST00000308874	PF07645	Calcium binding EGF domain
	ENST00000372476	PF07974	EGF-like domain
ENST00000331782	PF07645	Calcium binding EGF domain	

	ENST00000379446	PF00018	SH3 domain
	ENST00000216733	PF00018	SH3 domain
	ENST00000219069	PF01352	KRAB box
	ENST00000337673	PF01352	KRAB box
	ENST00000336034	PF01335	Death effector domain
	ENST00000268605	PF00619	Caspase recruitment domain
	ENST00000262320	PF00615	Regulator of G protein signaling domain
	ENST00000345122	PF00071	Ras family
	ENST00000345122	PF01846	FF domain
	ENST00000355619	PF00646	F-box domain
	ENST00000333602	PF00627	UBA/TS-N domain
	ENST00000373812	PF04146	YT521-B-like family
Other functions	ENST00000355810	PF01129	NAD:arginine ADP-ribosyltransferase
	ENST00000366899	PF00581	Rhodanese-like domain
	ENST00000305631	PF00487	Fatty acid desaturase
	ENST00000361971	PF01403	Sema domain
	ENST00000263574	PF00014	Kunitz/Bovine pancreatic
	ENST00000404535	PF01928	Trypsin inhibitor domain
	ENST00000361790	PF05624	CYTH domain
	ENST00000358602	PF00488	Lipolysis stimulated receptor (LSR)
	ENST00000264381	PF00135	MutS domain V
	ENST00000338660	PF00092	Carboxylesterase
	ENST00000258613	PF00090	von Willebrand factor

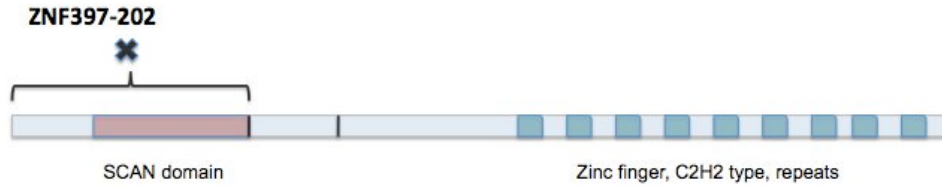


Figure 4.9: Example of a tissue-specific exon exclusion that removes the whole domain from a protein product. The ZNF397-202 isoform of Zinc finger protein 397 (ENST00000330501) encodes several DNA binding Zinc finger repeats and the protein interaction SCAN domain. Tissue-specific alternative splicing removes the exon that encodes the SCAN domain, thus possibly preventing the interactions that modulate action of this protein.

## 4.4 Discussion

### 4.4.1 Evolution and function of alternative splicing

Alternative splicing is considered to be a major source of functional diversity in animal proteins, particularly in mammals (Keren et al., 2010; Kondrashov and Koonin, 2003). Its major role is in increasing proteome diversity, but this mechanism also regulates transcript abundance through nonsense-mediated decay (Stamm et al., 2005). Data obtained by new sequencing technologies suggest that the degree of splicing in human genes is much higher than previously anticipated, with more than 95% of multiexon genes undergoing alternative splicing (Pan et al., 2008).

New alternative splice isoforms can be created by the insertion of new protein coding sequences that originated from noncoding sequences of introns (Kondrashov and Koonin, 2003). However, as discussed in Chapter 3, alternative splicing can also play an important role after exon shuffling, in particular after gene fusion, ensuring that ancestral protein products are expressed together with new protein isoforms. Finally, new splice isoforms can emerge after transition of a constitutive exon to an alternative exon (Lev-Maor et al., 2007). Interestingly, it has been found that the origin of an exon can influence how

frequently it is spliced into an mRNA (Modrek and Lee, 2003), with old exons more frequently being constitutive than younger exons.

Splicing can be regulated in a tissue- or developmental stage-specific manner. Such carefully regulated exons have been considered as a special class of exons in the previous studies as well, and specific regulation of an isoform was sometimes taken as support for its function (Lareau et al., 2004). In particular, tissue-specific exons were found to exhibit characteristics that can distinguish them from other types of exons. It was shown that tissue-specific exons tend to be highly conserved and modular – i.e. their length is often a multiple of three so inclusion or exclusion of these exons does not disrupt the translation of the rest of the protein (Xing and Lee, 2005). In this study, I observe that tissue-specific exons are enriched in functional disordered protein regions, which suggests that finely regulated expression of different splice isoforms of the same gene plays an important regulatory role.

Previous analyses of alternative splice isoforms of the same gene demonstrated that alternative splicing can determine the intracellular localization of a protein, enzymatic activity and stability, but also the posttranslational modifications and binding properties of a protein – including the binding of small ligands, nucleic acids and other proteins (Stamm et al., 2005). In line with this, it was suggested that alternative splicing bridges the gap between organism complexity and the number of genes in the organism not only by increasing the proteome size, but also by increasing the regulation and complexity of cellular networks (Lareau et al., 2004; Resch et al., 2004). Results from this study further emphasise the regulatory role of alternative splicing.

This study focused on alternatively spliced exons that encode functional residues which determine protein-protein interactions. However, alternative inclusion of other, even short, protein segments can have dramatic consequences for the overall protein function. A good illustration for this is the Piccollo protein (Garcia et al., 2004). This gene produces two protein isoforms that differ in nine residues. As a result of this, the shorter isoform has a stronger binding affinity for  $\text{Ca}^{2+}$ , but is also incapable of undergoing  $\text{Ca}^{2+}$ - dependent dimerization that normally occurs in a longer isoform. The structural study of this protein showed that this was a consequence of a large structural change induced by the omitted

short motif. Apart from causing a drastic change in protein structure, alternative splicing can also affect the connector region between the globular domains of a protein and in that way influences their orientation and recruitment of their binding partners. Additionally, splicing can also affect regions that determine ligand binding, which was not covered in this study. Hence, design of this study covers only a fraction of alternative splice events that can have important consequences for the overall protein function.

#### 4.4.2 Unstructured functional residues direct isoform-specific networks

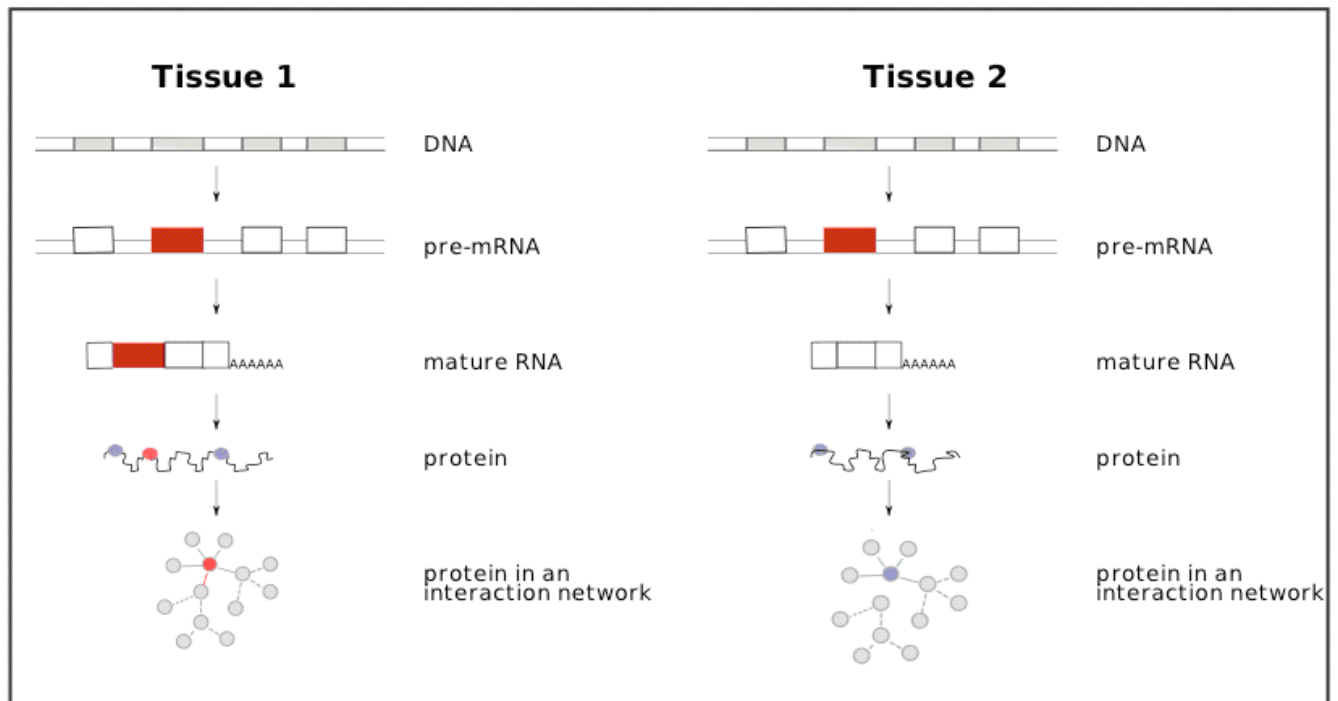
In this study, I observed a strong enrichment of tissue-specific exons in unstructured protein regions. Moreover, I also found that the disordered regions encoded by tissue-specific exons were likely to expose functional residues which determine binding interactions with other proteins. These binding interactions are determined by the exposed binding peptides and PTM sites.

Tissue-specific exons are overall more conserved than other exons; this has been reported before and is also confirmed by the results from this study (Xing and Lee, 2005) (Table 4.5). Interestingly, I observed that a large contribution to this high conservation of tissue-specific exons came from the exon regions that encode unstructured protein segments (Table 4.5). This can be explained either by the important function of the encoded disorder or by the conserved signals for exon splicing which overlap the residues that encode these disordered segments. However, predicted unstructured binding segments in tissue-specific exons are more conserved than predicted disordered regions, and are in fact more conserved than all other residues in these exons. Hence, this lends support to the claim that conserved disordered segments encoded by tissue-specific exons are indeed functional. Moreover, the high conservation of tissue-specific exons is likely to be also due to important binding motifs in these exons. Similarly, previous work has associated conserved disordered regions with DNA/RNA and protein binding functions (Chen et al., 2006).

Protein posttranslational modifications have emerged as a common regulatory switch in cell signalling networks. Moreover, it has been reported that

PTMs in general and protein phosphorylation in particular, tend to occur more frequently within intrinsically disordered protein regions than in ordered ones (Iakoucheva et al., 2004). Because of the flexibility of disorder regions, exposed PTM sites can easily, and specifically, interact with modifying enzymes. Hence, this also allows the introduced modifications to be readily reversible (Fuxreiter et al., 2007). Such modes of interactions are of significant benefit in regulation, signaling and network organization (Dunker et al., 2005). Hence, disordered regions are believed to be hot spots for regulation by posttranslational modification (Dyson and Wright, 2005). Here, I observe a strong correlation between the fraction of disorder and a fraction of PTM sites encoded by exons (Figure 4.5). Significant overrepresentation of PTM sites in tissue-specific exons provides further support for the role of these exons in cellular networks and the functional significance of disorder encoded by tissue-specific exons.

This study suggests an important interplay of finely regulated tissue-specific alternative splicing and disordered protein segments in cell signalling pathways. By this means, unstructured binding motifs can act as a mode of switching interaction partners and contributing to the re-wiring of signalling pathways. This implies an important role played by tissue-specific protein isoforms in specific protein interactions and consequentially their role in signalling and regulatory pathways. When the data for it become available, it will be interesting to see if alternative splicing specific for developmental and differentiation stages uses the same strategy as tissue-specific splicing. It has already been suggested that alternative splicing could determine the binding partners of proteins and consequentially direct cellular interaction networks (Resch et al., 2004; Stamm et al., 2005; Yura et al., 2006). This study confirms that this indeed is the case with tissue-specific exons and additionally, it explains the dominant mechanism for this. By exposing functional disordered segments, alternative splicing has an opportunity to re-wire signalling pathways dynamically at the post-transcriptional level (illustrated in Figure 4.11). Furthermore, by splicing in these regions, protein functional diversity can be achieved without compromising stability. Therefore, through alternative splicing of disordered regions, which act as mediators for interactions, protein networks can change depending on the context – e.g. tissue.



**Figure 4.11: Illustration of the predicted effect of tissue-specific alternative splicing of functional disordered residues.** Tissue-specific splicing and differential inclusion of exons frequently results in differential presence of a protein segment with specific binding motifs. Binding motifs are shown as blue (constitutively present) and red (tissue-specifically present) circles on proteins (wavy lines). The consequence of this is tissue-specific rewiring of protein networks. In the depicted network, proteins are shown as circles and connections with proteins that the protein shown above (a coloured circle in the network) directly interacts with are presented with continuous lines. Absence of a specific binding motif results in a loss of connection to one or more branches of a protein network.



### 4.4.3 Examples for the role of disordered protein segments in signal transduction

The role of disordered protein segments in mediating protein regulatory function is becoming increasingly appreciated. Another aspect of protein interactions that seems to be well explained by structural malleability of unstructured segments is the phenomenon of “moonlighting”, e.g. the ability of the same protein to have distinct binding partners and hence distinct functions (Tompa et al., 2005). The advantage of using disordered protein regions for mediating interactions lies in the fact that the same unstructured region can have overlapping interaction surfaces and can adopt different conformations after binding (Tompa, 2005). By this means, a protein can exert distinct functional effects, depending on the available binding partner. An example from the literature for the importance of tissue-specific splicing that I have described here is the p73 gene. This gene is a homologue of the p53 tumour suppressor gene and hence it is not unexpected that disordered regions would play an important role in its function. Namely, it is known that the N terminal region of the p53 protein plays an important regulatory role and is able to bind several protein partners (Chumakov, 2007), among which MDM2 (Kussie et al., 1996), RPA 70N (Bochkareva et al., 2005) and RNA polymerase II (Di Lello et al., 2006). Interestingly, this region has been reported to be completely disordered (Dawson et al., 2003) and spectrometric studies of the p53 protein showed that this protein was partially unstructured over its whole length (Bell et al., 2002). It was suggested that this could be an explanation for why it can interact with a multitude of protein partners.

Even though the assignment of genes to pathways they belong to is fairly incomplete (Wu et al., 2010), there is enough annotation of the genes with tissue-specific isoforms to observe here that there are pathways which are repeatedly connected with these genes. Genes with tissue-specific isoforms are significantly enriched in genes that are involved in the PDZ pathway (Table 4.8). Proteins with the PDZ domain are scaffold proteins that play an important role in signal transduction; in particular they help to anchor transmembrane proteins to the cytoskeleton and hold together signalling complexes (Ranganathan and

Ross, 1997). The PDZ proteins also play a crucial role in the organization of synaptic protein composition and structure. The PDZ domain has several modes of interaction (Figure 4.12a), but is specialized in binding short unstructured peptide motifs at the extreme C-termini of protein partners (Kim and Sheng, 2004; Nourry et al., 2003). An illustration for this is the interaction of the membrane-embedded voltage-activated potassium channel (Kv) with the PDZ containing scaffold protein PSD-95 (Magidovich et al., 2007). This interaction is mediated by the C-terminal segment of the Kv channel and is essential for the proper assembly and functioning of the synapse. Experiments involving C-terminal chains with different flexibility and length clearly demonstrated that intrinsic disorder in this segment modulates its interaction with the PDZ protein partner (Magidovich et al., 2007). The interaction, described as a “fishing rod mechanism”, is illustrated in Figure 4.12b. This experimental evidence highlights the importance of intrinsically disordered protein segments in complex processes of synapse assembly, maintenance and function. The ability of PDZ proteins to bind short extreme C-terminal sequences of their interaction partners offers an easy way for PDZ proteins to interact with target proteins without disrupting the overall structure and function of their protein partners, which are often membrane receptors bound to ligands (Hung 2002). Because of this, the PDZ proteins have a widespread role in synaptic signalling, in both the presynaptic and postsynaptic terminus. The role of protein disorder in the PDZ pathway is well established, and this study suggests that genes in this pathway can utilize tissue-specific expression of protein segments, which are likely to be disordered, as an extra mode of regulation. This is particularly interesting because the connection with alternative splicing suggests that some of the interactions in the pathway could be involving disordered regions present only in certain gene isoforms.

Genes with tissue-specific isoforms are enriched in genes from the PDZ pathway but are also reported to include genes from several pathways related to MAPK signalling (Table 4.8). As discussed in the introduction, this central signalling pathway can activate numerous cellular processes and represents a good hypothetical target for modulation of protein function through alternative inclusion of disordered binding residues. However, the role of functional

disordered residues has not been connected with this pathway so far. Nonetheless, the example of the MEK kinase shows disorder could be utilized in this pathway to direct specific signalling. The MEK kinase exists in two gene copies: MEK1 and MEK2. Sequences of their protein products are highly similar and their kinase domains essentially identical; they were initially even considered to be functionally redundant (Shaul and Seger, 2007). However, the proteins do differ in their N-termini and in the proline-rich inserts (residues phosphorylated by MAPK kinase kinases). As a result, each protein forms signalling complexes with different protein partners (Shaul and Seger, 2007). This has such strong implications that knockout of MEK1 causes an embryonic lethality in MEK1<sup>-/-</sup> mice whereas MEK2<sup>-/-</sup> mice are viable and fertile (Shaul and Seger, 2007). The analysis of the MEK1 and 2 protein sequences showed that their N-terminal regions are indeed unstructured (section 4.3.6).

Taken together, these examples illustrate the specific cases where protein disorder plays an important role and where finely regulated alternative splicing differentially exposes peptide motifs, which can be bound by other proteins, as a means to re-wire protein networks.

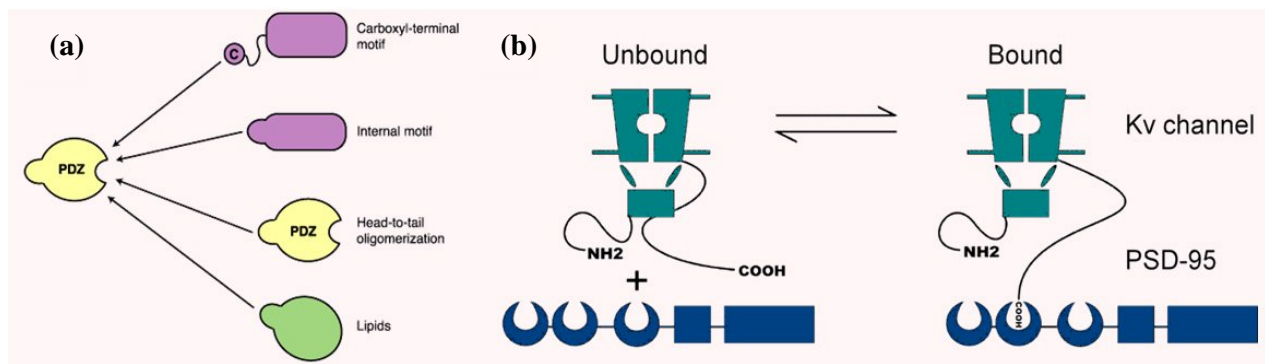


Figure 4.12: PDZ domain proteins play an important role in the targeting of proteins to specific membrane compartments and their assembly into supramolecular complexes. a) PDZ domains participate in at least four different classes of interaction: recognition of C-terminal motifs in peptides, recognition of internal motifs in peptides, PDZ-PDZ dimerization, and recognition of lipids. b) Interaction of a voltage-gated K<sup>+</sup> channel with a PSD-95 scaffold protein is an example of a fishing rod mechanism by which PDZ proteins interact with the unstructured C-termini of their protein partners. The moon-shape represent the PDZ domains of the PSD-95 protein. Figure (a) is adapted from (Nourry et al., 2003) and figure (b) from (Magidovich et al., 2007).

#### 4.4.4 Genes with tissue-specific isoforms and disease development

As discussed above, genes with tissue-specific isoforms are likely to play an important role in carefully regulated signalling pathways. Therefore, one can expect that mutations in these proteins are likely to have long-range consequences. In agreement with this, the set of tissue-specific genes is enriched with genes that were reported to cause embryonic lethality when mutated and are implicated in cancer development. Higher abundance of disordered regions among the cancer associated proteins has been suggested previously; 79% of human proteins associated with cancer have been classified as intrinsically unstructured, compared to 47% of all eukaryotic proteins in UniProtKB/Swiss-Prot (Iakoucheva et al., 2002). With regard to alternative splicing and cancer, it is known that mutations that affect splicing can have causal roles in cancer initiation and progression (Wang et al., 2002) and alternative splicing is in general frequently disrupted in cancer, though presumably mostly as a consequence of the overall instability in cancer cells (Venables, 2004). This study suggests a possible connection between the two and a role of isoforms with specific binding peptides in the pathways involved in cancer development.

The majority of protein domains, which are encoded by tissue-specific exons has a function related to binding (Table 4.9), emphasising that splicing can determine protein binding partners not only through alternative inclusion of unstructured binding motifs, but also by other means. RNA-binding proteins, which are essential for the production of alternative splice isoforms, could possibly work together with transcription factors in defining tissue-identity. The role of RNA-binding splicing factors in modulating the function of signalling proteins could be a part of the explanation for why these proteins are implicated in diseases that are connected with specific signalling pathways - both genetic disorders and cancer (Lukong et al., 2008).

By inclusion of disordered regions, functional capability of a single protein can expand depending on the context, space and time. When this process is related to disease development, it is an attractive target for drug application - especially if a drug, such as for example an antibody, can be made specific for

one isoform and not interfere with the function of other isoforms. However, in order to be able to interfere with this process, it is necessary first to understand it. More comprehensive studies of splicing and genomic architecture in an increasing number of species will surely play an important role in addressing this problem.

## 4.5 Bibliography

- Bell, S., Klein, C., Muller, L., Hansen, S., and Buchner, J. (2002). p53 contains large unstructured regions in its native state. *J Mol Biol* 322, 917-927.
- Bochkareva, E., Kaustov, L., Ayed, A., Yi, G.S., Lu, Y., Pineda-Lucena, A., Liao, J.C., Okorokov, A.L., Milner, J., Arrowsmith, C.H., *et al.* (2005). Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. *Proc Natl Acad Sci U S A* 102, 15412-15417.
- Bourdon, J.C. (2007). p53 and its isoforms in cancer. *Br J Cancer* 97, 277-282.
- Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., and Dunker, A.K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 55, 104-110.
- Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Blake, J.A. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* 36, D724-728.
- Chen, J.W., Romero, P., Uversky, V.N., and Dunker, A.K. (2006). Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J Proteome Res* 5, 888-898.
- Chumakov, P.M. (2007). Versatile functions of p53 protein in multicellular organisms. *Biochemistry (Mosc)* 72, 1399-1421.
- Consortium, U. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 37, D169-174.
- Dawson, R., Muller, L., Dehner, A., Klein, C., Kessler, H., and Buchner, J. (2003). The N-terminal domain of p53 is natively unfolded. *J Mol Biol* 332, 1131-1141.
- 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 Biol* 4, P3.
- Di Lello, P., Jenkins, L.M., Jones, T.N., Nguyen, B.D., Hara, T., Yamaguchi, H., Dikeakos, J.D., Appella, E., Legault, P., and Omichinski, J.G. (2006). Structure of the Tfb1/p53 complex: Insights into the interaction between the p62/Tfb1 subunit of TFIIF and the activation domain of p53. *Mol Cell* 22, 731-740.

- Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I., and Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* 5, 2985-2995.
- Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433-3434.
- Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* 272, 5129-5148.
- Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6, 197-208.
- Finn, R.D., Tate, J., Mistry, J., Coggill, 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.
- Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., and Stratton, M.R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet Chapter 10*, Unit 10 11.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. *Nat Rev Cancer* 4, 177-183.
- Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23, 950-956.
- Garcia, J., Gerber, S.H., Sugita, S., Sudhof, T.C., and Rizo, J. (2004). A conformational switch in the Piccolo C2A domain regulated by alternative splicing. *Nat Struct Mol Biol* 11, 45-53.
- Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8, R250.
- Gsponer, J., and Babu, M.M. (2009). The rules of disorder or why disorder rules. *Prog Biophys Mol Biol* 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* 322, 1365-1368.

- Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., and Iakoucheva, L.M. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* 2, e100.
- Holt, L.J., Tuch, B.B., Villen, J., Johnson, A.D., Gygi, S.P., and Morgan, D.O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325, 1682-1686.
- Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol* 4, R70.
- Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., *et al.* (2009). Ensembl 2009. *Nucleic Acids Res* 37, D690-697.
- Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., and Dunker, A.K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323, 573-584.
- Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32, 1037-1049.
- Ishimoto, O., Kawahara, C., Enjo, K., Obinata, M., Nukiwa, T., and Ikawa, S. (2002). Possible oncogenic potential of DeltaNp73: a newly identified isoform of human p73. *Cancer Res* 62, 636-641.
- Jin, P., Fu, G.K., Wilson, A.D., Yang, J., Chien, D., Hawkins, P.R., Au-Young, J., and Stuve, L.L. (2004). PCR isolation and cloning of novel splice variant mRNAs from known drug target genes. *Genomics* 83, 566-571.
- Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11, 345-355.
- Kim, E., and Sheng, M. (2004). PDZ domain proteins of synapses. *Nat Rev Neurosci* 5, 771-781.
- Kondrashov, F.A., and Koonin, E.V. (2003). Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet* 19, 115-119.
- Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., *et al.* (2009).



- ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics* 93, 213-220.
- Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing. *Trends Genet* 19, 124-128.
- Kussie, P.H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A.J., and Pavletich, N.P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 274, 948-953.
- Lareau, L.F., Green, R.E., Bhatnagar, R.S., and Brenner, S.E. (2004). The evolving roles of alternative splicing. *Curr Opin Struct Biol* 14, 273-282.
- Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., Leibman-Barak, S., Pupko, T., and Ast, G. (2007). The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet* 3, e203.
- Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends Genet* 24, 416-425.
- Magidovich, E., Orr, I., Fass, D., Abdu, U., and Yifrach, O. (2007). Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K<sup>+</sup> channel modulates its interaction with scaffold proteins. *Proc Natl Acad Sci U S A* 104, 13022-13027.
- Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* 5, e1000376.
- Modrek, B., and Lee, C.J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet* 34, 177-180.
- Nourry, C., Grant, S.G., and Borg, J.P. (2003). PDZ domain proteins: plug and play! *Sci STKE* 2003, RE7.
- Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635-648.
- Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40, 1413-1415.

- Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7, 208.
- Ranganathan, R., and Ross, E.M. (1997). PDZ domain proteins: scaffolds for signaling complexes. *Curr Biol* 7, R770-773.
- Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., and Lee, C. (2004). Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res* 3, 76-83.
- Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., *et al.* (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A* 103, 8390-8395.
- Shaul, Y.D., Gibor, G., Plotnikov, A., and Seger, R. (2009). Specific phosphorylation and activation of ERK1c by MEK1b: a unique route in the ERK cascade. *Genes Dev* 23, 1779-1790.
- Shaul, Y.D., and Seger, R. (2007). The MEK/ERK cascade: from signaling specificity to diverse functions. *Biochim Biophys Acta* 1773, 1213-1226.
- Shimizu, K., and Toh, H. (2009). Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. *J Mol Biol* 392, 1253-1265.
- Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005). Function of alternative splicing. *Gene* 344, 1-20.
- Taylor, J., Schenck, I., Blankenberg, D., and Nekrutenko, A. (2007). Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics Chapter* 10, Unit 10 15.
- Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579, 3346-3354.
- Tompa, P., Szasz, C., and Buday, L. (2005). Structural disorder throws new light on moonlighting. *Trends Biochem Sci* 30, 484-489.
- Tress, M.L., Bodenmiller, B., Aebersold, R., and Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol* 9, R162.
- Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Olason, P.I., Albrecht, M., Hegyi, H., Giorgetti, A., *et al.* (2007). The

- implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci U S A* *104*, 5495-5500.
- Venables, J.P. (2004). Aberrant and alternative splicing in cancer. *Cancer Res* *64*, 7647-7654.
- Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* *456*, 470-476.
- Wang, K., Geren, L., Zhen, Y., Ma, L., Ferguson-Miller, S., Durham, B., and Millett, F. (2002). Mutants of the CuA site in cytochrome c oxidase of *Rhodospirillum rubrum*: II. Rapid kinetic analysis of electron transfer. *Biochemistry* *41*, 2298-2304.
- Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* *293*, 321-331.
- Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* *11*, R53.
- Xing, Y., and Lee, C.J. (2005). Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet* *1*, e34.
- Yura, K., Shionyu, M., Hagino, K., Hijikata, A., Hirashima, Y., Nakahara, T., Eguchi, T., Shinoda, K., Yamaguchi, A., Takahashi, K., *et al.* (2006). Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene* *380*, 63-71.

## Chapter 5

### Concluding remarks

An organism's phenotype is primarily determined by the proteins its genome encodes. A crucial biological question is how protein repertoires have expanded in more complex organisms and how regulation of more complex proteomes is achieved. In my thesis, I addressed this problem by studying two means for the increase of proteome size: creation of novel proteins during evolution and alternative inclusion of functional modules in different isoforms of the same gene. My approach here was to look at the architecture of functional elements in proteins, investigate mechanisms that include or exclude these elements from the proteins, and consequences this has for the overall protein function.

In the first part of the thesis, I used animal gene phylogenies to investigate trends that shaped the evolution of protein domain architectures. Protein domains form the basic unit of protein functional and structural complexity. Furthermore, proteins with novel domain combinations had a major role in evolutionary innovation. Thus, formation of novel proteins through domain shuffling is a crucial aspect of animal evolution. The results of my study confirmed previous observations that changes in protein domain composition occur preferentially at the protein termini. Additionally, the study suggested that the same trend was present after both inferred gains and losses of single copy domains. Since different mechanisms can underlie insertions and deletions of single copy domains, it is possible that the observed pattern is not only shaped

by the acting mechanisms, but also by the selective pressure which strongly disfavors changes in the middle of proteins. Changes in the middle are more likely to disrupt the ancestral protein structure and hence only a small fraction of these are expected to get fixed in a population. A bias for the changes to occur at the termini is not as strong for duplications and deletions of domains in repeats. Nevertheless, different mechanisms and evolutionary forces can contribute to the evolution of domain repeats. The design of this study allowed me to distinguish changes in domain architecture that followed gene duplication from those that occurred after speciation. Interestingly, the same positional pattern of changes was observed for both types of events. Hence, this implies that changes in an individual protein are modeled similarly after both types of evolutionary events. However, the frequency of changes was two times higher after gene duplications, which indicated that the pressure to preserve the ancestral domain composition is relieved after a gene is present in two copies.

Even though the position of a domain gain or loss in a protein can discriminate between certain mechanisms that cause the changes, it cannot clearly specify the underlying mechanism. In the second part of this thesis, I focused on the investigation of the evidence for the mechanisms that were driving emergence of more complex domain architectures during evolution of animal gene families. In prokaryotes, new domains are predominantly acquired through fusions of adjacent genes. However, the relative contributions of the different molecular mechanisms that cause domain gains in animals were unknown. A crucial step here was to obtain a set of high confidence domain gains, and to relate these gains to the changes in the gene structures. For this, I again relied on the phylogenetic data that described the evolution of animal gene families. Results of this study showed that the major mechanism for gains of new domains in metazoan proteins was gene fusion through joining of exons from adjacent genes, possibly mediated by non-allelic homologous recombination. Two other mechanisms that were previously suggested to have an important role in the evolution of metazoans - retroposition and insertion of exons into ancestral introns through intronic recombination - appear to be only minor contributors to overall domain gains. Interestingly, the results of this study also suggested exon extensions through inclusion of previously non-coding regions as

an important mechanism for addition of disordered segments to proteins. In the case of confident domain gains, I observed that gene duplication preceded domain gain in at least 80% of the gain events. The interplay of gene duplication and domain gain demonstrates an important mechanism for fast neofunctionalisation of genes. Interestingly, the gained domains are frequently involved in protein interactions. Hence, this illustrates a fundamental connection between the evolution of proteome diversity and regulation of more complex cellular networks.

In addition to evolutionary changes in the architectures of protein functional elements, novel protein products can also be created through alternative inclusion of exons from the same gene. By this means, the gene's function can adapt to different cellular contexts. In the final part of this thesis, I investigated how finely regulated alternative inclusion of tissue-specific exons modifies protein function. I observed a strong trend for tissue-specific exons to encode the segments enriched in intrinsically disordered regions. I found that these alternatively spliced protein segments were also significantly enriched in binding peptides and post-translationally modified sites. Functional relevance of the observed phenomenon was further indicated by significant evolutionary conservation of the tissue-specific disordered regions and predicted binding peptides. By alternatively splicing functional disordered segments, an individual gene can achieve functional versatility without compromising the structural stability of its protein products. In addition, different protein isoforms of the same gene can be used in different cellular networks. This could also be one of the mechanisms for the regulation of tissue-specific signalling pathways. It is a frequent phenomenon that the same gene takes part in cellular pathways that have different, sometimes even opposing, outcomes. Intriguingly, mechanisms that ensure the specificity of the transmitted signals are still unclear. This research suggests that it is possible that finely regulated alternative splicing of functional disordered protein segments can assist in attaining this specificity. Since the mechanisms for regulation of signalling specificity are frequently disrupted in cancer and other diseases, it is important to understand the contribution of this process in the regulation of signalling cascades. In conclusion, extension of proteins with novel interaction domains and alternative

inclusion of disordered binding segments demonstrate two different effective means for the increase of proteome size and a level of proteome regulation.

The work in this thesis emphasises the impact that inclusion or exclusion of protein functional elements has on its role in an organism. Both changes on the gene level and changes on the transcript level can modify the architecture of functional elements in the final protein product. Improved characterization and coverage of proteins with these elements – protein domains, binding peptides and post-translationally modified sites – can help in better understanding of the effect that these changes can have on protein function, and in understanding how this drives protein evolution and adaptation to different tissues and cellular contexts. Additionally, I expect that application of new technologies for sequencing not just genomes, but also transcriptomes in different organisms and tissues will improve our understanding of the areas that I address in this thesis. Identifying transcripts that are specific for an organism or a tissue is a good starting point for describing proteins that define tissues, or organism phenotypes, and can provide more complete datasets for similar studies.

A problem that I find particularly interesting is the effect of a change in the number of short repeated domains, since these are crucial for cellular interactions. A change in the number of domains in a repeat can change protein's affinity for the binding partners and hence affect the whole cellular interaction network. To adequately address this issue, it would be first necessary to have high quality domain annotations. One means to increase the quality of these annotations is to lower the threshold for assignment of repeated domains - in particular after the first domain from a repeat has already been assigned to a protein, and in order to avoid false assignments - to require that a short repeated domain, when annotated, is present in a protein with its whole length. Finally, to better understand how a change in the number of domains in a repeat, or the presence or absence of other functional elements in proteins, influences protein functions, it would be valuable to have good quality functional annotations for different protein homologues and isoforms of the same gene. Relating a certain type of a change in the architecture of protein functional elements to the overall change in protein function would allow us to better understand the consequences that each change can introduce in less-studied proteins.

# Appendices



## Appendix A

Table Appendix A.1: Possible false positive Pfam assignments in the TreeFam proteins. A list of Pfam domains that only a single gene in a gene family is annotated with, and that are not predicted with a high E-value nor cover a high fraction of a domain model.

TreeFam family	Pfam domain	Domain name	Fraction of a model covered	E-value
TF101021	PF01154	HMG_CoA_synt_N	0.14	0.0011
TF101181	PF01576	Myosin_tail_1	0.08	0.00018
TF101220	PF00621	RhoGEF	0.26	2.60E-05
TF102023	PF08092	Toxin_22	0.3	0.0081
TF105126	PF08609	Fes1	0.17	0.00086
TF105285	PF00021	UPAR_LY6	0.15	0.00022
TF105388	PF00580	UvrD-helicase	0.08	0.00011
TF105664	PF07602	DUF1565	0.07	1.30E-05
TF105993	PF08624	CRC_subunit	0.25	0.0001
TF106336	PF05693	Glycogen_syn	0.07	0.0004
TF106337	PF03488	Ins_beta	0.27	1.30E-05
TF300142	PF01370	Epimerase	0.27	0.00018
TF300253	PF00125	Histone	0.28	0.0002
TF300491	PF00128	Alpha-amylase	0.16	0.00017
TF300506	PF08001	CMV_US	0.12	0.00031
TF300523	PF02689	Herpes_Helicase	0.04	3.00E-05
TF300533	PF02672	CP12	0.21	5.30E-05
TF300647	PF08764	Coagulase	0.05	0.02
TF300805	PF05585	DUF1758	0.06	0.00043
TF312998	PF00398	RrnaAD	0.13	0.00046
TF313187	PF02790	COX2_TM	0.2	0.017
TF313234	PF01757	Acyl_transf_3	0.28	2.10E-05
TF313377	PF07732	Cu-oxidase_3	0.27	0.017
TF313568	PF08634	Pet127	0.05	0.0052
TF313594	PF08443	RimK	0.19	5.00E-05
TF313654	PF02932	Neur_chan_memb	0.14	4.10E-05
TF313802	PF06807	Clp1	0.1	0.0013
TF313930	PF04258	Peptidase_A22B	0.08	1.70E-05
TF313947	PF01271	Granin	0.07	0.011
TF314126	PF05904	DUF863	0.02	3.50E-05
TF314165	PF00136	DNA_pol_B	0.09	4.50E-05
TF314440	PF08529	NusA_N	0.18	0.00091
TF314441	PF06127	DUF962	0.01	6.80E-05
TF314495	PF08401	DUF1738	0.15	0.0001
TF314521	PF00650	CRAL_TRIO	0.15	0.0054
TF314774	PF01595	DUF21	0.13	2.10E-05
TF315186	PF06282	DUF1036	0.18	3.60E-05

TF315189	PF00851	Peptidase_C6	0.05	0.00029
TF315227	PF05511	ATP-synt_F6	0.29	5.10E-05
TF315263	PF00600	Flu_NS1	0.12	0.0098
TF315272	PF08320	PIG-X	0.08	0.00088
TF315302	PF08637	NCA2	0.08	0.0026
TF315363	PF00636	Ribonuclease_3	0.14	0.00011
TF315367	PF02752	Arrestin_C	0.26	1.90E-05
TF315472	PF01546	Peptidase_M20	0.29	0.00044
TF315592	PF00775	Dioxygenase_C	0.01	0.0011
TF315712	PF03564	DUF1759	0.26	3.10E-05
TF315897	PF01537	Herpes_glycop_D	0.11	2.40E-05
TF316533	PF07462	MSP1_C	0.04	0.00028
TF316780	PF03238	ESAG1	0.06	0.0053
TF316929	PF02932	Neur_chan_memb	0.13	0.00026
TF317006	PF06650	DUF1162	0.12	1.90E-05
TF317757	PF02093	Gag_p30	0.15	0.0043
TF317925	PF00878	CIMR	0.12	0.00022
TF318379	PF01030	Recep_L_domain	0.26	4.90E-05
TF318668	PF00001	7tm_1	0.01	2.80E-05
TF318706	PF08719	DUF1768	0.15	1.20E-05
TF319588	PF00650	CRAL_TRIO	0.15	0.0054
TF319633	PF07714	Pkinase_Tyr	0.01	2.20E-05
TF319951	PF01461	7tm_4	0.18	7.00E-05
TF321275	PF02682	AHS1	0.08	0.00019
TF321359	PF00106	adh_short	0.29	0.00044
TF321457	PF00443	UCH	0.25	1.20E-05
TF321796	PF05473	Herpes_UL45	0.17	2.00E-05
TF322230	PF02752	Arrestin_C	0.26	1.90E-05
TF323518	PF07933	DUF1681	0.13	0.00012
TF323731	PF00650	CRAL_TRIO	0.17	0.002
TF323819	PF01237	Oxysterol_BP	0.18	0.00014
TF323965	PF00094	VWD	0.15	9.60E-05
TF323987	PF00147	Fibrinogen_C	0.16	0.00011
TF324336	PF08389	Xpo1	0.26	2.50E-05
TF324755	PF09409	PUB	0.26	0.00067
TF324880	PF00261	Tropomyosin	0.07	0.00018
TF325457	PF03052	Adeno_52K	0.16	0.001
TF325523	PF08583	UPF0287	0.21	0.00053
TF326264	PF00168	C2	0.23	0.00015
TF326378	PF02355	SecD_SecE	0.14	0.00029
TF326897	PF00122	E1-E2_ATPase	0.23	2.20E-05
TF328040	PF01613	Flavin_Reduct	0.14	0.00012
TF329290	PF07933	DUF1681	0.13	0.001
TF329430	PF00168	C2	0.25	8.10E-05
TF329606	PF00447	HSF_DNA-bind	0.22	1.30E-05
TF329606	PF01579	DUF19	0.12	0.021
TF329710	PF02250	Orthopox_35kD	0.12	0.00046
TF330156	PF07732	Cu-oxidase_3	0.27	2.90E-05
TF330183	PF05806	Noggin	0.08	4.60E-05
TF330319	PF03401	Bug	0.18	0.00035
TF330845	PF00012	HSP70	0.11	0.00038
TF331115	PF00555	Endotoxin_M	0.29	0.0033
TF331158	PF05642	Sporozoite_P67	0.04	7.20E-05
TF331282	PF00386	C1q	0.26	9.10E-05

TF331344	PF00822	PMP22_Claudin	0.12	2.90E-05
TF331400	PF02790	COX2_TM	0.21	0.0077
TF331842	PF01429	MBD	0.25	0.00033
TF332204	PF05253	UPF0224	0.19	2.60E-05
TF332241	PF00059	Lectin_C	0.29	2.00E-05
TF332364	PF02373	JmjC	0.27	2.60E-05
TF332426	PF07798	DUF1640	0.24	1.80E-05
TF332497	PF08124	Lyase_8_N	0.06	0.0006
TF332538	PF05579	Peptidase_S32	0.05	7.90E-05
TF332659	PF08562	Crisp	0.21	0.0011
TF332845	PF04266	ASCH	0.18	0.00011
TF333186	PF09451	ATG27	0.09	5.20E-05
TF333434	PF00836	Stathmin	0.24	0.0017
TF333463	PF00081	Sod_Fe_N	0.21	0.00082
TF333601	PF03255	ACCA	0.2	0.001
TF335097	PF01271	Granin	0.05	0.0082
TF335573	PF02622	DUF179	0.12	0.0003
TF335835	PF01068	DNA_ligase_A_M	0.24	0.00025
TF338389	PF02825	WWE	0.3	0.0016
TF338479	PF00100	Zona_pellucida	0.23	0.0003
TF339541	PF06039	Mqo	0.05	0.00014
TF339848	PF01579	DUF19	0.15	0.012
TF340612	PF03154	Atrophin-1	0.04	0.00079
TF341730	PF01347	Vitellogenin_N	0.14	0.0002

## Appendix B

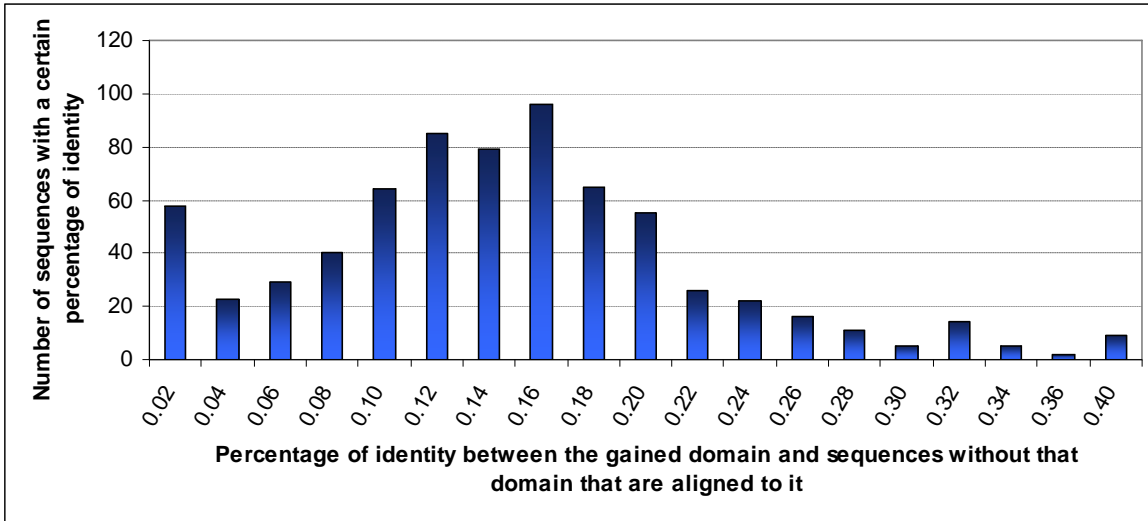


Figure Appendix B.1: Distribution of the percentages of identity between the inferred gained domain and the most similar sequence in the same gene family that does not have that domain assigned. The set of domain gains that is shown in the graph was filtered to include only internal gains and gains that have a descendant with the gained domain in at least one genome of a better quality. Two sequences in the same gene family are aligned either because of the shared ancestry, or because multiple alignment algorithms (MUSCLE in this case, <http://www.drive5.com/muscle>) over-align similar regions in proteins, even when they are not evolutionarily related. Both of these instances are likely to be present in the regions where inferred gained domains are aligned to sequences of other proteins in the same gene family. The peak at 0.16 could be explained with addition of values from these two scenarios.

Table Appendix B.2: High confidence domain gain events. Information about descendants of the gain event is shown only for the gains in the human lineage.

TreeFam family	Pfam domain	Representative transcript	Descendants
TF340491	PF02518	ENST00000275580	Primates
TF331377	PF04680	ENST00000290291	Primates
TF352220	PF05033,PF00856	ENST00000307483	Primates
TF331083	CL0074	ENST00000338965	Primates
TF342157	PF04698	ENST00000354668	Primates
TF328297	CL0219,PF02023	ENST00000357581	Primates
TF340395	CL0159	ENST00000359050	Primates
TF314793	PF00271,CL0008	ENST00000370424	Primates
TF351422	PF10409	ENST00000381866	Primates
TF105356	CL0023	ENST00000194097	Mammals
TF335271	CL0041	ENST00000254691	Mammals
TF328011	PF02023	ENST00000259883	Mammals
TF337552	PF00096	ENST00000262637	Mammals
TF328424	PF05386	ENST00000262715	Mammals
TF337951	PF00147	ENST00000301455	Mammals
TF300253	PF03002	ENST00000320498	Mammals
TF350810	PF01352	ENST00000338637	Mammals
TF338854	PF01352	ENST00000344099	Mammals
TF331962	PF00612	ENST00000366709	Mammals
TF338165	PF04711	ENST00000367990	Mammals
TF336000	PF08065	ENST00000368654	Mammals
TF330114	CL0175	ENST00000373330	Mammals
TF333425	PF04593	ENST00000388827	Mammals
TF325887	PF10522	ENST00000394516	Mammals
TF105660	PF08062	ENST00000399466	Mammals
TF330855	PF03523	ENST00000262101	Mammals
TF334740	CL0006,PF00621	ENST00000296794	Mammals
TF329807	PF06049	ENST00000367797	Mammals
TF324004	PF02008	ENST00000373644	Mammals
TF317779	PF09307	ENST00000009530	Vertebrates
TF326567	CL0003	ENST00000046794	Vertebrates
TF325130	PF00023	ENST00000160373	Vertebrates
TF106374	CL0172	ENST00000199447	Vertebrates
TF105392	CL0159,PF03160	ENST00000200181	Vertebrates
TF325426	PF00632	ENST00000206595	Vertebrates
TF319848	PF01033	ENST00000229003	Vertebrates
TF106352	PF00023	ENST00000230792	Vertebrates
TF313285	PF01759,PF01821	ENST00000245907	Vertebrates
TF329176	CL0081	ENST00000249910	Vertebrates
TF330078	PF10393	ENST00000255132	Vertebrates
TF320327	PF04812	ENST00000257555	Vertebrates
TF320327	PF04813	ENST00000257555	Vertebrates
TF330114	PF05485	ENST00000260045	Vertebrates
TF331062	CL0072	ENST00000260283	Vertebrates
TF313938	CL0154	ENST00000260983	Vertebrates
TF316484	PF01463,CL0022,PF01822	ENST00000262304	Vertebrates
TF316148	PF03815	ENST00000262424	Vertebrates
TF312824	PF00612	ENST00000262457	Vertebrates
TF316113	CL0003	ENST00000262878	Vertebrates
TF317402	CL0159	ENST00000263798	Vertebrates

TF317511	PF00017	ENST00000263915	Vertebrates
TF331945	PF07525	ENST00000264607	Vertebrates
TF316876	PF00093	ENST00000264895	Vertebrates
TF316876	PF03160	ENST00000264895	Vertebrates
TF324610	PF02732	ENST00000267430	Vertebrates
TF351678	PF01392	ENST00000273857	Vertebrates
TF314731	PF10565	ENST00000279593	Vertebrates
TF313240	PF10606	ENST00000282753	Vertebrates
TF316380	CL0011	ENST00000283296	Vertebrates
TF316380	PF01390	ENST00000283296	Vertebrates
TF351678	CL0202	ENST00000284885	Vertebrates
TF329158	CL0023	ENST00000285928	Vertebrates
TF314232	PF00569,CL0220	ENST00000288642	Vertebrates
TF316484	PF02010	ENST00000289672	Vertebrates
TF332664	PF07776	ENST00000289816	Vertebrates
TF313285	CL0005	ENST00000291440	Vertebrates
TF336193	PF01342,PF03172	ENST00000291582	Vertebrates
TF323966	CL0214	ENST00000294383	Vertebrates
TF317402	PF01403	ENST00000296474	Vertebrates
TF329295	CL0124	ENST00000296498	Vertebrates
TF329059	CL0001	ENST00000296575	Vertebrates
TF331157	CL0041	ENST00000297350	Vertebrates
TF312852	CL0219	ENST00000298139	Vertebrates
TF323475	CL0003	ENST00000298229	Vertebrates
TF323480	CL0005	ENST00000302495	Vertebrates
TF331319	PF01822,CL0164	ENST00000303746	Vertebrates
TF106506	PF00023	ENST00000303941	Vertebrates
TF106401	PF00249	ENST00000310806	Vertebrates
TF327329	PF00051,PF09396	ENST00000311907	Vertebrates
TF314204	PF02816	ENST00000313478	Vertebrates
TF324155	PF00023	ENST00000313581	Vertebrates
TF315996	CL0006	ENST00000314276	Vertebrates
TF316105	CL0188	ENST00000317133	Vertebrates
TF317614	PF06959	ENST00000317905	Vertebrates
TF315956	PF05485	ENST00000321679	Vertebrates
TF317659	PF01391	ENST00000322313	Vertebrates
TF329915	PF00040	ENST00000323926	Vertebrates
TF106510	PF02161	ENST00000325455	Vertebrates
TF313103	PF07941	ENST00000328224	Vertebrates
TF318980	PF02165	ENST00000332351	Vertebrates
TF333138	PF01391	ENST00000333570	Vertebrates
TF329287	PF00642	ENST00000333834	Vertebrates
TF317921	PF00023	ENST00000340022	Vertebrates
TF316214	PF04621	ENST00000343495	Vertebrates
TF105669	PF00458	ENST00000344102	Vertebrates
TF318080	CL0016	ENST00000344204	Vertebrates
TF315606	CL0041	ENST00000344227	Vertebrates
TF329345	CL0010	ENST00000344936	Vertebrates
TF321873	CL0056	ENST00000355044	Vertebrates
TF326161	PF01284	ENST00000355237	Vertebrates
TF300189	PF10574	ENST00000357484	Vertebrates
TF330032	PF01033	ENST00000357639	Vertebrates
TF300851	PF00642	ENST00000357720	Vertebrates
TF328589	PF09303	ENST00000358316	Vertebrates
TF323607	PF06462	ENST00000359520	Vertebrates
TF323475	PF00017	ENST00000359570	Vertebrates
TF315592	PF01392	ENST00000360986	Vertebrates
TF331681	PF00057	ENST00000361205	Vertebrates
TF326495	PF06663	ENST00000367213	Vertebrates
TF315841	PF02205	ENST00000367288	Vertebrates

TF334159	PF05177	ENST00000367856	Vertebrates
TF315806	CL0123	ENST00000368474	Vertebrates
TF314133	CL0003	ENST00000369075	Vertebrates
TF329606	PF03509	ENST00000369235	Vertebrates
TF316297	PF06839	ENST00000369466	Vertebrates
TF316833	PF06484	ENST00000371130	Vertebrates
TF313103	PF03521	ENST00000371741	Vertebrates
TF101106	PF10487	ENST00000372577	Vertebrates
TF331727	PF05604	ENST00000372970	Vertebrates
TF312900	CL0202	ENST00000373187	Vertebrates
TF330498	CL0154	ENST00000373209	Vertebrates
TF330345	CL0011	ENST00000373401	Vertebrates
TF320194	CL0196	ENST00000373638	Vertebrates
TF300648	CL0172	ENST00000375663	Vertebrates
TF300648	PF00043	ENST00000375663	Vertebrates
TF313965	PF00084	ENST00000377034	Vertebrates
TF331310	CL0033	ENST00000377674	Vertebrates
TF106001	PF02344,PF01056	ENST00000377970	Vertebrates
TF313698	PF03700	ENST00000380285	Vertebrates
TF315592	CL0202	ENST00000380605	Vertebrates
TF324293	CL0154	ENST00000380868	Vertebrates
TF316876	CL0056	ENST00000380881	Vertebrates
TF332820	PF08365	ENST00000381389	Vertebrates
TF329720	CL0084,PF00533	ENST00000381989	Vertebrates
TF323983	CL0179	ENST00000383733	Vertebrates
TF105391	CL0128	ENST00000389202	Vertebrates
TF317067	CL0266	ENST00000389247	Vertebrates
TF316056	PF09004	ENST00000389568	Vertebrates
TF318198	CL0188	ENST00000389821	Vertebrates
TF106341	PF00010	ENST00000389936	Vertebrates
TF330156	PF03815	ENST00000392504	Vertebrates
TF331707	CL0219,PF09091	ENST00000392723	Vertebrates
TF331055	CL0010	ENST00000393398	Vertebrates
TF336041	CL0001	ENST00000394980	Vertebrates
TF337303	PF00435	ENST00000395209	Vertebrates
TF314963	CL0208	ENST00000396197	Vertebrates
TF317532	CL0011	ENST00000396906	Vertebrates
TF317402	PF01833,PF01437	ENST00000397752	Vertebrates
TF106276	PF08959	ENST00000398892	Vertebrates
TF331207	PF00014	ENST00000399429	Vertebrates
TF106451	PF07452	ENST00000204604	Bilateria
TF314081	CL0033	ENST00000215739	Bilateria
TF313754	PF00805	ENST00000221200	Bilateria
TF331485	PF00988,CL0014	ENST00000233072	Bilateria
TF323999	PF00773	ENST00000252889	Bilateria
TF323502	PF02185	ENST00000254260	Bilateria
TF324918	PF00057,CL0186	ENST00000260197	Bilateria
TF313551	PF08912	ENST00000261535	Bilateria
TF313326	CL0190	ENST00000261875	Bilateria
TF323159	CL0020	ENST00000263635	Bilateria
TF351276	CL0072	ENST00000264042	Bilateria
TF354308	CL0221	ENST00000278279	Bilateria
TF315363	PF00611	ENST00000281092	Bilateria
TF324744	PF00642	ENST00000295373	Bilateria
TF315892	CL0010	ENST00000295713	Bilateria
TF318935	PF02218	ENST00000301843	Bilateria
TF315897	PF03765	ENST00000306726	Bilateria
TF323280	PF00630	ENST00000323468	Bilateria
TF318014	PF00412	ENST00000336180	Bilateria
TF101179	PF09465	ENST00000338179	Bilateria

TF315363	CL0266	ENST00000348343	Bilateralialia
TF323999	PF07145	ENST00000358691	Bilateralialia
TF323312	PF00641	ENST00000359653	Bilateralialia
TF324164	CL0223	ENST00000369443	Bilateralialia
TF326321	PF01424,CL0196	ENST00000371527	Bilateralialia
TF324293	CL0266,PF00621	ENST00000380868	Bilateralialia
TF323674	PF02825	ENST00000389044	Bilateralialia
TF314351	CL0126	ENST00000061240	AllAnimals
TF330032	CL0263	ENST00000075322	AllAnimals
TF329240	CL0200	ENST00000202017	AllAnimals
TF313988	PF04707	ENST00000251170	AllAnimals
TF314316	PF01463,CL0022	ENST00000252804	AllAnimals
TF335359	PF06009	ENST00000252999	AllAnimals
TF314796	CL0041	ENST00000261600	AllAnimals
TF317296	CL0266	ENST00000261752	AllAnimals
TF320906	PF00787	ENST00000262211	AllAnimals
TF313191	PF08403	ENST00000262461	AllAnimals
TF105399	PF06466	ENST00000263754	AllAnimals
TF313184	PF00595	ENST00000264431	AllAnimals
TF323502	CL0031	ENST00000265562	AllAnimals
TF317067	CL0006	ENST00000268676	AllAnimals
TF314219	PF02809	ENST00000289528	AllAnimals
TF102004	CL0072	ENST00000295797	AllAnimals
TF314470	CL0186	ENST00000298125	AllAnimals
TF316118	PF00439	ENST00000302054	AllAnimals
TF314638	CL0183	ENST00000310298	AllAnimals
TF300359	CL0220	ENST00000310454	AllAnimals
TF312822	CL0271	ENST00000311630	AllAnimals
TF105056	CL0137	ENST00000313698	AllAnimals
TF314677	PF09141	ENST00000314888	AllAnimals
TF314748	CL0154	ENST00000324068	AllAnimals
TF319230	PF00023	ENST00000332509	AllAnimals
TF106173	PF02148	ENST00000334136	AllAnimals
TF312960	CL0010	ENST00000338257	AllAnimals
TF313629	CL0266	ENST00000339416	AllAnimals
TF316643	PF00373	ENST00000340930	AllAnimals
TF318080	CL0011	ENST00000344204	AllAnimals
TF316643	PF03623	ENST00000346049	AllAnimals
TF105282	PF08070	ENST00000348049	AllAnimals
TF314159	PF06311	ENST00000355058	AllAnimals
TF106448	CL0114	ENST00000357008	AllAnimals
TF106151	CL0196	ENST00000358896	AllAnimals
TF313758	PF00880	ENST00000359988	AllAnimals
TF351123	CL0159	ENST00000360304	AllAnimals
TF323658	PF00397	ENST00000361125	AllAnimals
TF105224	CL0186	ENST00000361961	AllAnimals
TF314076	CL0186	ENST00000367097	AllAnimals
TF354311	CL0221	ENST00000367122	AllAnimals
TF300807	PF02225	ENST00000367512	AllAnimals
TF320809	CL0010	ENST00000369405	AllAnimals
TF314566	PF09162	ENST00000372788	AllAnimals
TF317034	PF00620	ENST00000373026	AllAnimals
TF314897	PF01585	ENST00000373451	AllAnimals
TF323767	CL0003	ENST00000373886	AllAnimals
TF319104	PF00880	ENST00000377187	AllAnimals
TF323577	PF00784	ENST00000377307	AllAnimals
TF314263	CL0016	ENST00000378168	AllAnimals
TF324293	CL0010	ENST00000380868	AllAnimals
TF300785	PF07533	ENST00000382194	AllAnimals
TF102004	CL0266	ENST00000392038	AllAnimals



TF314028	PF00355	ENST00000399167	AllAnimals
TF323674	PF06701	ENST00000399332	AllAnimals
TF332135	PF06046	ENSMUST00000011407	
TF316484	PF02140	ENSMUST00000040422	
TF316155	PF02178	ENSMUST00000040802	
TF328297	PF06747	ENSMUST00000041466	
TF335390	PF00096	ENSMUST00000051869	
TF329295	PF00100	ENSMUST00000084509	
TF344032	CL0016	ENSMUST00000086209	
TF352132	PF02415	ENSMUST00000087258	
TF335097	PF00530	ENSMUST00000090986	
TF343969	CL0072	ENSMUST00000096028	
TF350794	PF01352	ENSMUST00000098508	
TF327726	PF08742	ENSMUST00000098633	
TF106451	PF08742,PF01826,PF00094	ENSMUST00000101614	
TF313537	CL0164	ENSMUST00000102891	
TF331090	CL0188	ENSMUST00000106224	
TF313147	CL0202	ENSMUST00000106949	
TF334740	CL0266	ENSMUST00000109426	
TF332078	PF03172	ENSMUST00000113392	
TF317514	CL0069,CL0011	ENSRNOT00000011676	
TF101514	PF08155	ENSRNOT00000012798	
TF319471	PF00096	ENSRNOT00000034133	
TF337163	PF08384	ENSRNOT00000041557	
TF335163	PF03501	ENSRNOT00000043986	
TF314473	CL0010	ENSXETT00000002556	
TF336376	PF10479	ENSXETT00000010407	
TF313664	PF00628,CL0008	ENSXETT00000017293	
TF352568	PF01759	ENSXETT00000037556	
TF327588	CL0291	ENSXETT00000041950	
TF343001	PF06512	ENSXETT00000045061	
TF343800	CL0266	ENSXETT00000049369	
TF343807	PF02135	ENSXETT00000049701	
TF330284	CL0164	ENSXETT00000055961	
TF316425	CL0004	ENSGALT00000003763	
TF330943	CL0102	ENSGALT00000012528	
TF343232	PF00612	ENSGALT00000017818	
TF331401	CL0066	ENSGALT00000036204	
TF326300	PF02181	ENSDART00000002526	
TF326300	PF02205	ENSDART00000002526	
TF342779	PF00681	ENSDART00000026448	
TF333311	PF08344	ENSDART00000045905	
TF329914	CL0184	ENSDART00000054641	
TF335519	PF08441	ENSDART00000055263	
TF318964	CL0266	ENSDART00000075811	
TF343508	CL0016	ENSDART00000076763	
TF316498	PF05485	ENSDART00000078494	
TF300180	CL0202	ENSDART00000078606	
TF329039	CL0188	ENSDART00000080545	
TF350019	PF00260	ENSDART00000081597	
TF106435	PF00748	ENSDART00000081614	
TF315837	PF02188,CL0072	ENSDART00000084559	
TF330777	PF01049	ENSDART00000086138	
TF315536	CL0272	ENSDART00000087610	
TF315645	CL0001	ENSDART00000097691	
TF332213	CL0229	ENSDART00000098581	
TF351676	PF01033	ENSDART00000104096	
TF343963	PF07500	ENSDART00000036529	
TF335838	PF07776	CG10431-RA	
TF343858	PF05030	CG10555-RA	

TF327367	PF05267	CG10912-RA
TF326895	CL0229	CG10916-RA
TF325393	PF02757	CG11066-RB
TF313668	PF04568	CG11079-RA
TF343304	CL0081	CG13598-RA
TF332191	CL0155	CG13676-RA
TF326889	PF00079	CG14470-RA
TF319090	CL0155	CG14608-RA
TF344100	PF06818	CG15365-RA
TF325916	CL0056	CG15378-RA
TF329913	PF08742,PF00094	CG15671-RA
TF350188	PF02757	CG15731-RA
TF324584	PF00631	CG15844-RA
TF351124	CL0011	CG16974-RA
TF316403	PF01049	CG17941-RA
TF321823	PF00014	CG18296-RA
TF312905	CL0220	CG31216-RA
TF323648	CL0004	CG32226-RA
TF343869	PF03128	CG32580-RA
TF343188	CL0155	CG32656-RA
TF351975	PF00650	CG32697-RA
TF319052	PF05444	CG34040-RA
TF319243	PF00412	CG4656-RA
TF343781	PF01753	CG4877-RA
TF315391	PF07776	CG5034-RA
TF313817	CL0229	CG5071-RB
TF316895	PF00023	CG5424-RB
TF343612	CL0155	CG5756-RA
TF327546	PF05267	CG5765-RA
TF313415	PF10545	CG6279-RA
TF105127	CL0031	CG7042-RA
TF313080	CL0265	CG7067-RA
TF336220	PF00569,CL0220	CG8529-RC
TF317532	PF01753,CL0049	CG8569-RA
TF322044	PF00628	CG8677-RA
TF316872	CL0004	CG9138-RA
TF314883	PF07773	CG9227-RA
TF317819	PF09607	CG9653-RA
TF105292	CL0220	CG9847-RA
TF326676	CL0126	CG9850-RB

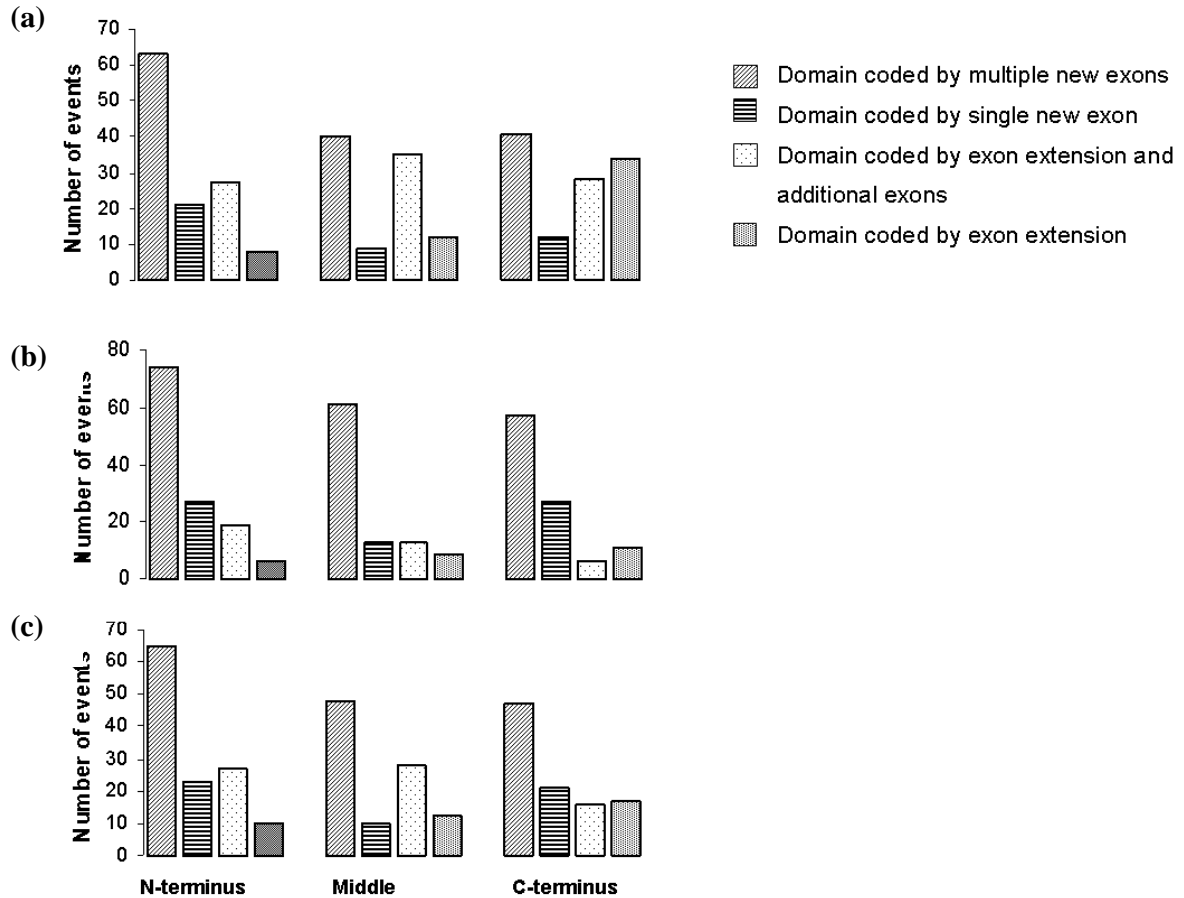


Figure Appendix B.3: Distribution of high confidence domain gain events according to the position of domain insertion and number of exons gained according to three different classification methods. (a) Method 1 was used for classification of high confidence domain gain events in Chapter 3. (b) Method 2 classifies a domain gain as an exon extension if there are at least 20 amino acids towards either of the exon borders and at least 30% of these are identical to a sequence in the alignment that does not contain the gained domain. (c) Method 3 classifies a domain gain as exon extension if there are at least 15 amino acids towards either of the exon borders and at least 25% of these are identical to a sequence in the alignment that does not contain the gained domain. Methods 2 and 3 classify each domain gain as a gain at the termini if towards the termini there are at least 80% unaligned residues or there are less than 10% identical residues in any of the sequences without the gained domain. There were seven and six domain gain events with ambiguous positions obtained by Methods 2 and 3, respectively which are not included in the Figure.

Table Appendix B.4: Domains that are gained by insertion of new exons(s) into the introns of ancestral genes. Phases of introns that surround the exons coding for the gained domains are shown for each gain event. In two cases (marked with \* next to domain name) introns surrounding domains did not have symmetrical phases, however additional exons appeared to have been gained together with the one(s) coding for these domains and phases of introns surrounding all inserted exons were symmetrical. It is also noted whether the gained domain(s) is/are coded by single or multiple new exons.

TreeFam family	Domain gained	Phase of 5' intron	Phase of 3' intron	Is single exon coding for the gained domain
TF336041	CL0001	1	1	Yes
TF335097	PF00530	1	1	Yes
TF331962	PF00612*	1	1	Yes
TF313965	PF00084	1	1	Yes
TF351678	CL0202	1	1	No
TF330156	PF03815	1	1	No
TF329915	PF00040	1	1	No
TF325130	PF00023	1	1	No
TF324293	CL0010	1	1	No
TF323674	PF06701	1	1	No
TF321873	CL0056	1	1	No
TF318080	CL0011	1	1	No
TF317532	CL0011	1	1	No
TF317402	CL0159	1	1	No
TF316484	PF02140	1	1	No
TF316380	CL0011	1	1	No
TF315592	PF01392	1	1	No
TF315592	CL0202	1	1	No
TF313537	CL0164	1	1	No
TF105391	CL0128	1	1	No
TF331319	PF01822,CL0164	1	1	No
TF324293	CL0266,PF00621	1	1	No
TF314677	PF09141	0	0	No
TF314133	CL0003	0	0	No
TF314081	CL0033*	0	0	No
TF313551	PF08912	0	0	No
TF300785	PF07533	0	0	No
TF106435	PF00748	0	0	No
TF325887	PF10522	0	1	Yes
TF324610	PF02732	0	1	Yes
TF323999	PF00773	1	2	Yes
TF322044	PF00628	0	1	Yes
TF315892	CL0010	1	2	Yes
TF354311	CL0221	2	1	No
TF350794	PF01352	0	2	No
TF335359	PF06009	1	0	No
TF331062	CL0072	2	1	No
TF329158	CL0023	2	1	No
TF319848	PF01033	1	2	No
TF319230	PF00023	2	1	No
TF317921	PF00023	2	0	No
TF317614	PF06959	0	2	No
TF316118	PF00439	0	1	No

TF314638	CL0183	2	0	No
TF313938	CL0154	0	2	No
TF313629	CL0266	0	1	No
TF312900	CL0202	1	0	No
TF106448	CL0114	0	1	No
TF327329	PF00051,PF09396	1	2	No

## Appendix C

Table Appendix C.1: Predicted binding sites and annotated PTM sites in the set of tissue-specific exons compared to the sets of cassette and constitutive exons. Column headed 'N<sub>+</sub>' shows the number of exons with the examined characteristic, 'N<sub>-</sub>' of those without it, and column headed 'Fraction<sub>+</sub>' shows a fraction of exons with the examined characteristic. PTM sites were taken from the UniProtKB/Swiss-Prot database. The P-value shows the results of the comparison with the set of Tissue-specific exons, and is obtained with the Chi-square test.

Analysis	Set of exons	N <sub>+</sub>	N <sub>-</sub>	N <sub>total</sub>	Fraction <sub>+</sub>	P-value
<b>ANCHOR</b>	Tissue-specific	410	1,016	1,426	0.288	N/A
	Cassette	8,821	40,203	49,024	0.180	P<2.2x10 <sup>-16</sup>
	Constitutive	27,374	122,564	149,938	0.183	P<2.2x10 <sup>-16</sup>
<b>PTM sites</b>	Tissue-specific	119	798	917	0.130	N/A
	Cassette	1,521	20,272	21,793	0.070	P=9.9x10 <sup>-12</sup>
	Constitutive	7,671	85,360	93,031	0.082	P=3.2x10 <sup>-7</sup>

Table Appendix C.2: All BioCarta pathways, and clusters of BioCarta pathways that a set of genes with tissue-specific isoforms is enriched in. EASE P-values represent modified Fisher exact P-values. Column 'Benjamini' shows P-values after applying the Benjamini correction for multiple tests.

Pathway	EASE P-value	Benjamini P-value
<i>Enriched individual pathways:</i>		
<b>PDZ pathway:</b> Synaptic Proteins at the Synaptic Junction	2.3x10 <sup>-5</sup>	7x10 <sup>-3</sup>
<b>IntegrinPathway:</b> Integrin Signaling Pathway	0.09	0.89
<b>MapkPathway:</b> MAPKinase Signaling Pathway	0.06	0.90
<b>HifPathway:</b> Hypoxia-Inducible Factor in the Cardiovascular System	0.09	0.90
<b>Pitx2Pathway:</b> Multi-step Regulation of Transcription by Pitx2	0.09	0.90
<b>p35alzheimersPathway:</b> Deregulation of CDK5 in Alzheimers Disease	0.08	0.90
<b>BiopeptidesPathway:</b> Bioactive Peptide Induced Signaling Pathway	0.06	0.92
<b>VegfPathway:</b> VEGF, Hypoxia, and Angiogenesis	0.08	0.92
<b>Her2Pathway:</b> Role of ERBB2 in Signal Transduction and Oncology	0.05	0.93
<b>CaCaMPathway:</b> Ca <sup>++</sup> / Calmodulin-dependent Protein Kinase Activation	0.06	0.93
<b>NdkDynaminPathway:</b> Endocytotic role of NDK, Phosphins and Dynamin	0.04	0.94
<b>At1rPathway:</b> Angiotensin II mediated activation of JNK Pathway via Pyk2 dependent signaling	0.04	0.97
<b>RhoPathway:</b> Rho cell motility signaling pathway	0.02	0.98

<b>ArapPathway:</b> ADP-Ribosylation Factor	0.04	0.98
<i>Enriched clusters of pathways with similar gene members:</i>		
<b>Cluster I</b>		
<b>MapkPathway:</b> MAPKinase Signaling Pathway	0.06	0.90
<b>p38mapkPathway:</b> p38 MAPK Signaling Pathway	0.26	0.98
<b>ErkPathway:</b> Erk1/Erk2 Mapk Signaling pathway	0.35	0.99
<b>Cluster II</b>		
<b>At1rPathway:</b> Angiotensin II mediated activation of JNK Pathway via Pyk2 dependent signaling	0.04	0.97
<b>BiopeptidesPathway:</b> Bioactive Peptide Induced Signaling Pathway	0.06	0.92
<b>IntegrinPathway:</b> Integrin Signaling Pathway	0.09	0.89
<b>pyk2Pathway:</b> Links between Pyk2 and Map Kinases	0.11	0.91
<b>Fcer1Pathway:</b> Fc Epsilon Receptor I Signaling in Mast Cells	0.18	0.97
<b>Cxcr4Pathway:</b> CXCR4 Signaling Pathway	0.18	0.96
<b>EcmPathway:</b> Erk and PI-3 Kinase Are Necessary for Collagen Binding in Corneal Epithelia	0.18	0.96
<b>BcrPathway:</b> BCR Signaling Pathway	0.60	1.00
<b>MetPathway:</b> Signaling of Hepatocyte Growth Factor Receptor	0.69	1.00
<b>TcrPathway:</b> T Cell Receptor Signaling Pathway	0.82	1.00