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 repertoir of gene function, it is necessarry 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).

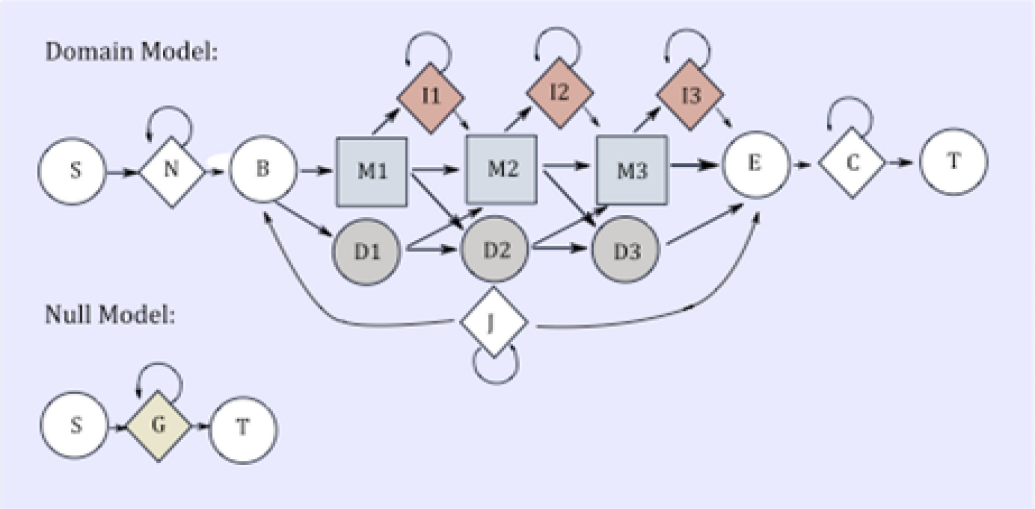

Figure 1.1: Diagram of profile hidden Markov model. States shown as squares or diamonds emit symbols, while those shown as circles do not. Each match state Mi corresponds to a column in a multiple alignment which emits over a distribution of amino acids. Insert states Ii allow for the segments of query 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 posttranslationfally 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.

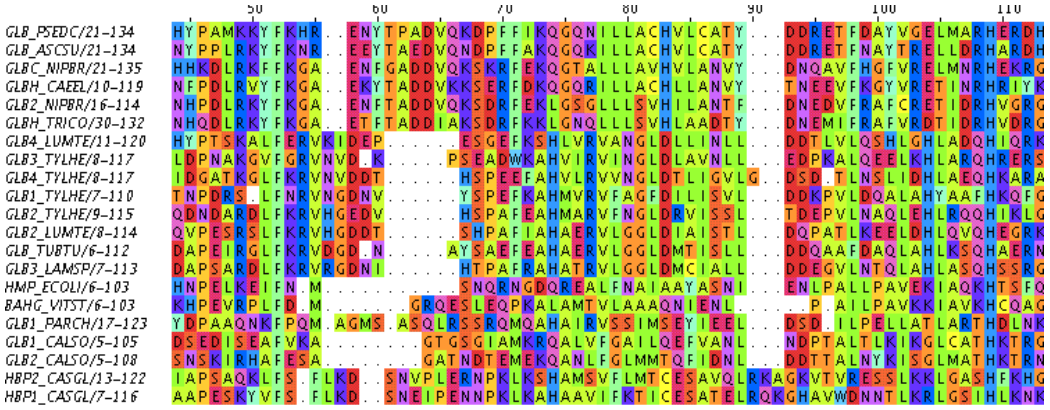

Figure 1.2: An example of a seed multiple alignment for the Pfam Globin family (Pfam accesion: PF00042). The seed alignment is used to build an HMM model of a family, which is the used for identifying the same domain in other proteins.

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 (Gsponer 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 measurments (Gsponer 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 easilfy 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 (Gsponer 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.

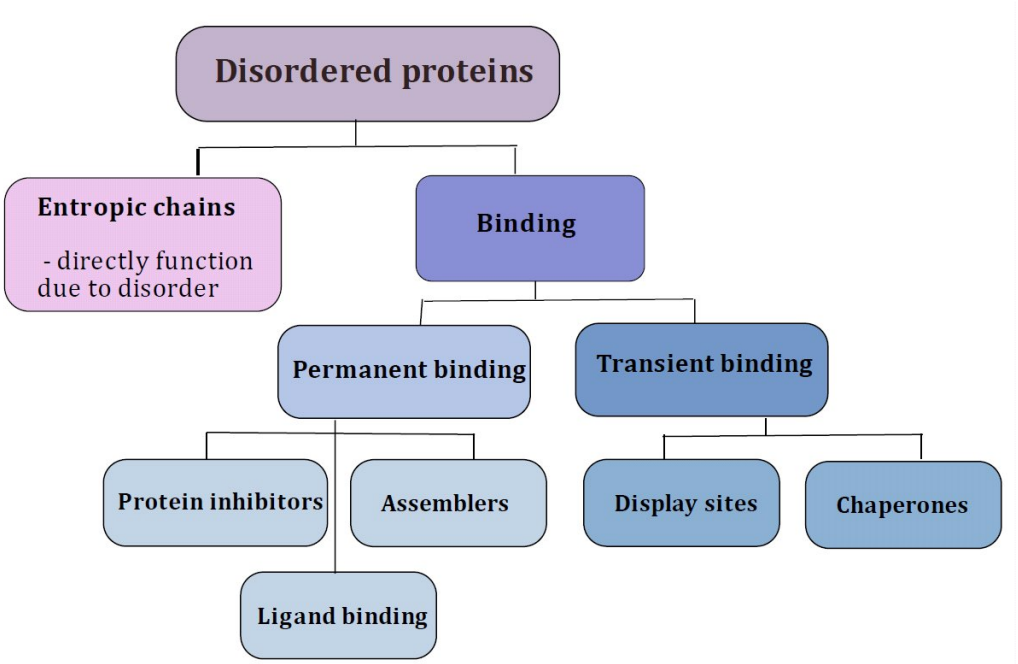

Figure 1.3: Functional classification of disordered proteins. Examples of disordered proteins from each category are described in the text. Illustration is adapted from Tompa (2005).

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 expection 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.

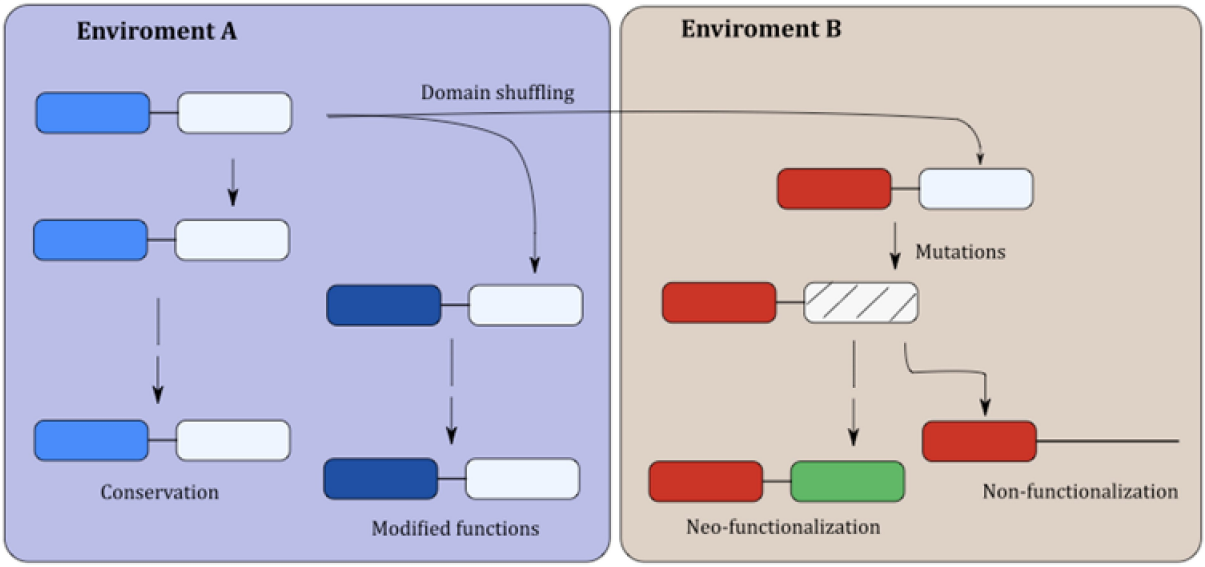

Figure 1.4: Domain shuffling and domain evolution. When domain shuffling changes the enviroment of a domain, the domain is likely to experience more radical changes in sequence and function. The domain enviroment is defined by the subcellular localization and interaction partners of a domain. The figure 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 slowely than if the domain is attached to a protein that operates in a different cellular compartment and/or has different proten 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 *Drosphila* 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 (Ciccarelli 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 (Ciccarelli 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).

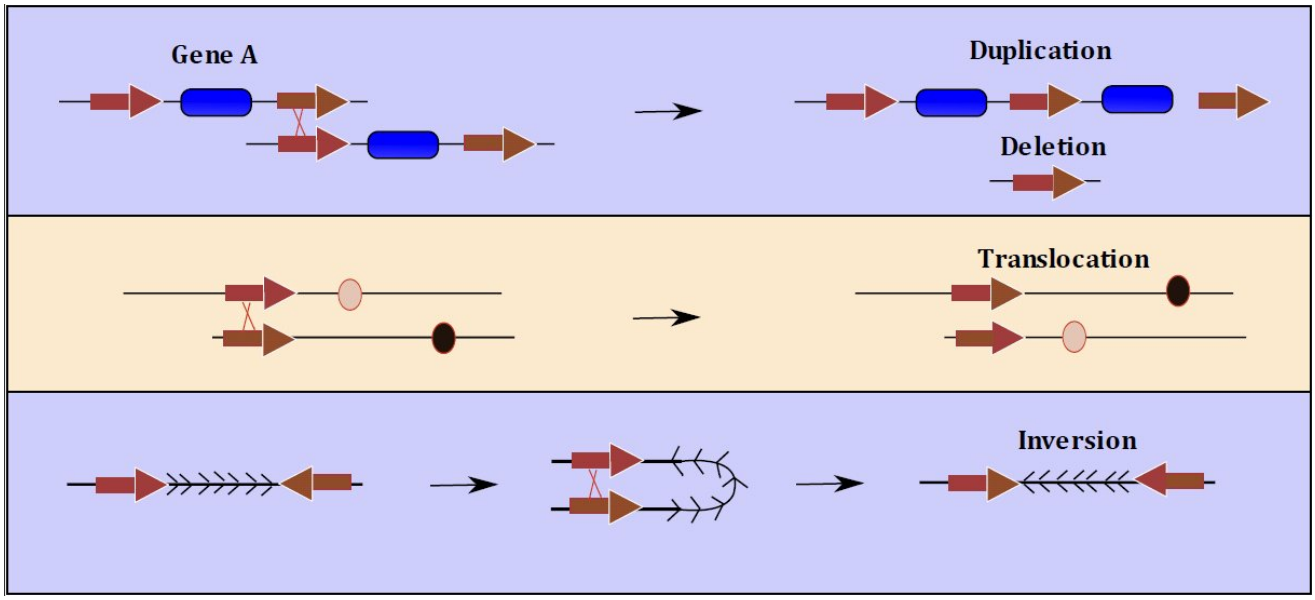

Figure 1.5: Possible effects of Non-allelic homologous recombination (NAHR) on genome evolution. NAHR between two highly similar segments in the genome can cause different types of rearrangements, depending on the location and orientation of these segments. Thus, NAHR between adjacent duplicated sequences 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.

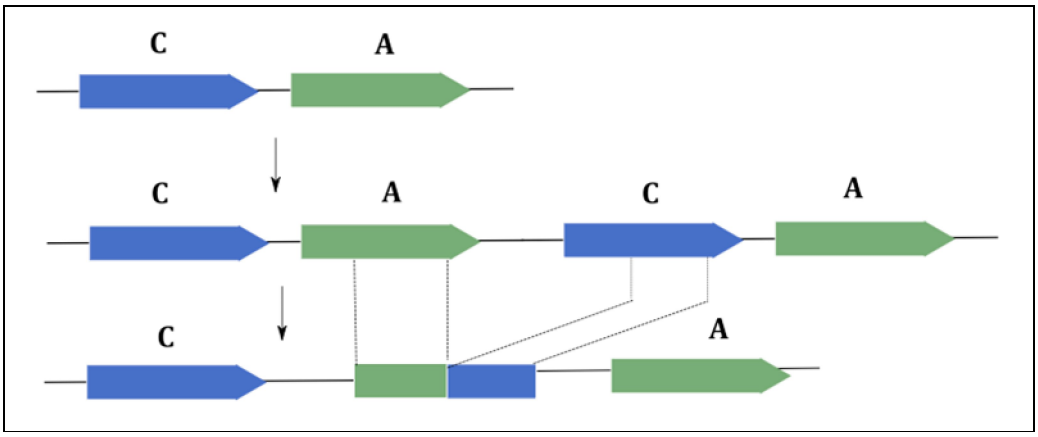

Figure 1.6: Example of a chimeric gene formed by gene fusion. The model is a simplified scenario of the evolution of the *Sdic* gene. Steps in the evolution of this gene include tandem duplication of neighbouring genes named C and A. This is followed with the deletion of parts of genes A and C as well 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 absorbtion 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 archea 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 archea species, as for example, *Methanocaldococcis 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 β-globin cluster in humans. One gene from this cluster is expressed specifically in embryos, another in foetuses 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 orthologus 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 Moult, 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 exluded 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).

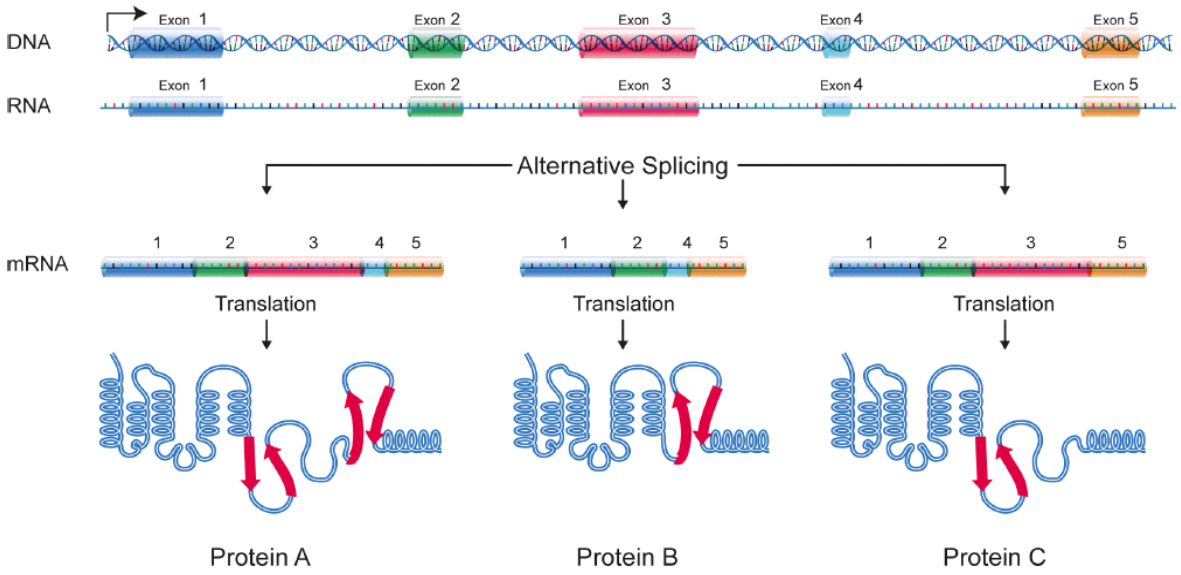

Figure 1.7: Alternative splicing increases the diversity of proteome. Alternative inclusion of exons 3 and 4 in this example can change the structure and function of the resulting protein products. Figure is taken from: http://upload.wikimedia.org/wikipedia/commons/0/0a/DNA_alternative_splicing.gif

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., Coggill, 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 pathwaysrepairing 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 Moult, J. (2009). Structural implication of splicing stochastics. 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.