# 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.