Chapter 4

Protein products of tissue-specific alternative splicing

4.1 Introduction

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

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

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

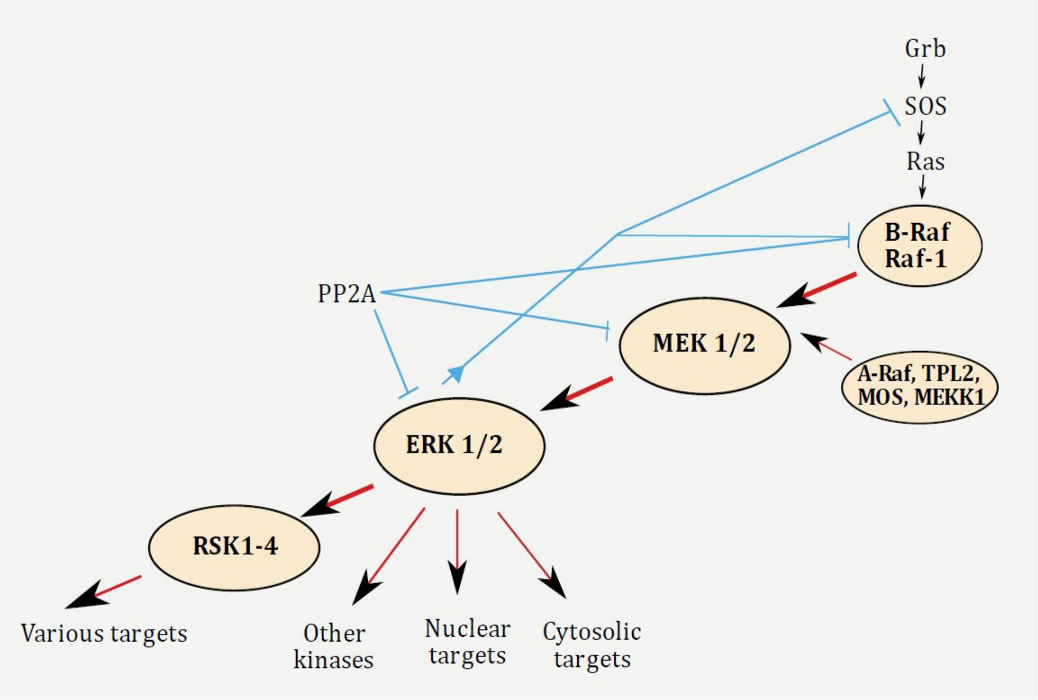

Figure 4.1: Schematic representation of the ERK signaling cascade. The bold arrows show the main pathway upon growth factor activation. Red arrows show activatory phosphorylation events, green accessory phosphorylation and blue inhibitory phosphorylation and dephosphorylation. The illustration is adapted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

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

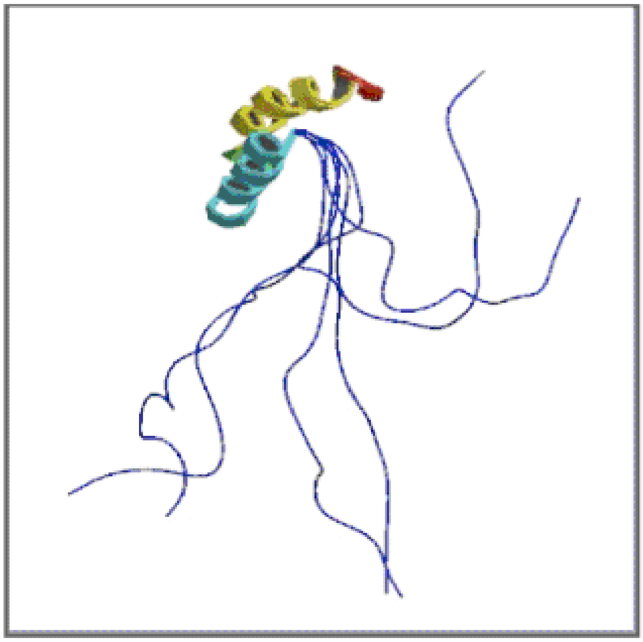

Figure 4.2: The structure of a human mitochondrial protein apoCox17 illustrates a protein with both ordered and disordered regions. The change of the structure colour from blue to red indicates direction of the sequence from N- to C-terminus. The positions of amino acids at the disordered N-terminus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

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

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

4.2 Methods

4.2.1 Sets of tissue-specific, cassette and constitutive exons

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

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

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

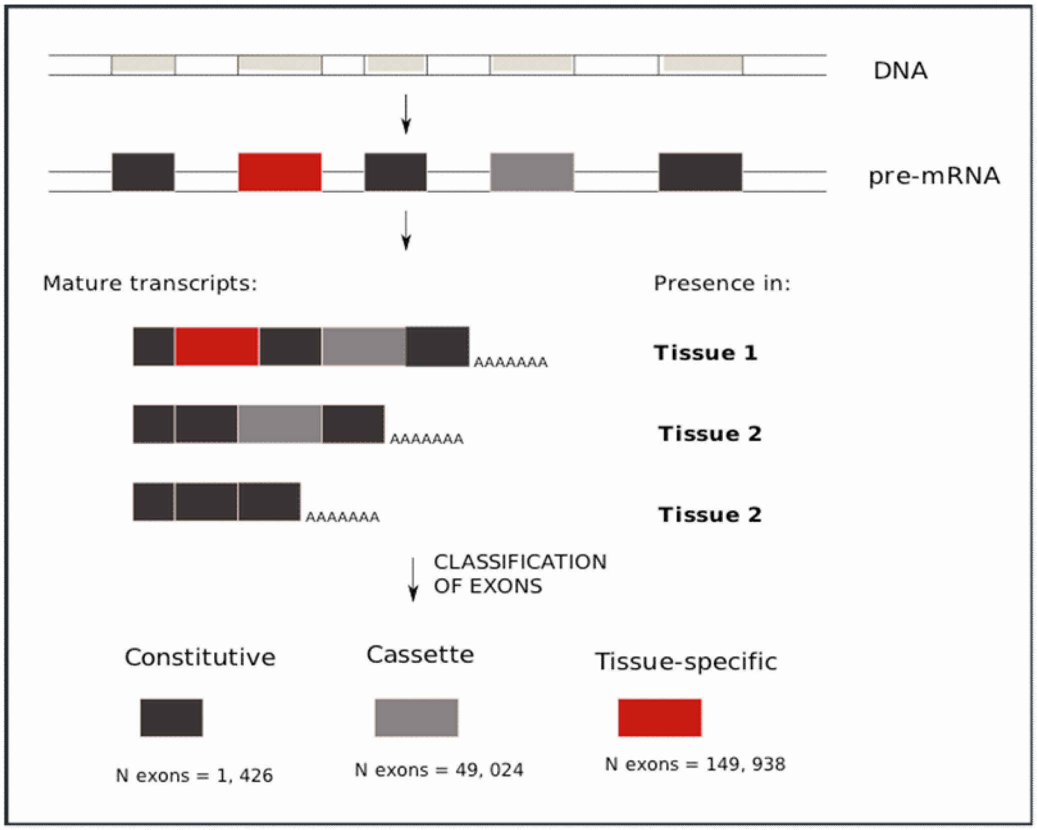

Figure 4.3: Scheme of exon classification. Tissue-specific exons are obtained from the study by Wang *et al.*, while sets of cassette and constitutive exons are made by classifying Ensembl coding exons according to this scheme.

4.2.3 Prediction of disordered protein residues

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

4.2.4 Prediction of functional residues

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

4.2.5 Conservation of exons in the three different datasets

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

4.2.6 Significance of observed trends

To test whether the differences in the fractions of disordered residues, predicted binding motifs, annotated PTM sites and experimentally predicted phosphosites in the three sets of exons were significant Chi-square tests were applied by using the R software. Significance of exon and peptide conservation in the tissue-specific set compared to two other sets, as well as conservation of peptide versus all other residues in the tissue-specific set, were tested with the Mann-Whitney tes (Wilcox test in the R software). The Mann-Whitney test was applied because the distribution of exon conservation values did not follow the normal distribution ($P<2.2x10^{-16}$, Shapiro-Wilk test for the distribution of values for tissue-specific exons). Test sets of cassette and constitutive exons with the same average length as in the set of tissue-specific exons were composed and fractions of predicted binding motifs and annotated PTM sites were calculated. The significance in the difference of fractions of the predicted functional residues was tested with the Chi-square test, again using the R software.

4.2.7 Comparison of MEK1 and MEK2 protein sequences

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

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

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

4.2.9 Disorder signatures in the protein products of the p73 gene

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

4.3 Results

4.3.1 Sets of exons with different expression profiles

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

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

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

GO term descripttion	GO term ID	EASE P-value	Benjamini P-value
Protein binding	0005515	5.4x10^{-16}	1.6x10^{-12}
Cytoskeletal protein binding	0008092	9.7x10^{-13}	1.4x10^{-9}
Actin binding	0003779	4.3x10^{-10}	4.1x10^{-07}
Binding	0005488	1.1x10^{-05}	7.7x10^{-03}
Catalytic activity	0003824	1.8x10^{-05}	1.0x10^{-02}
Transferase activity	0016740	2.1x10^{-5}	1.0x10^{-2}
Transferase activity, transferring phosphorus-containing groups	0016772	3.2x10^{-05}	1.3x10^{-02}
Kinase activity	0016301	4.4x10^{-5}	1.6x10^{-2}
Protein serine/threonine kinase activity	0004674	5.3x10^{-05}	1.7x10^{-02}
Enzyme binding	0019899	1.1x10^{-04}	3.2x10^{-02}
Nucleotide binding	0000166	1.3x10^{-04}	3.4x10^{-02}
Ras GTPase binding	0017016	1.9x10^{-04}	4.3x10^{-02}

4.3.2 Tissue-specific exons are enriched in disordered residues

I compared the fractions of disordered residues in the three sets of exons with different expression profiles. Figure 4.2 shows a protein which contains both ordered and disordered regions. Disordered regions were identified using the IUPred software (Dosztanyi et al., 2005), which predicts unstructured protein regions in the segments with biased amino acid composition, such as those enriched in polar or charged residues, which do not allow formation of sufficient stabilizing interactions. I found that both sets of alternatively spliced exons - the set of tissue-specific and the set of cassette exons - were enriched with exons encoding disordered amino acids, when compared to the set of constitutive exons (Figure 4.4). The fraction of exons coding for unstructured protein regions was the highest for the tissue-specific exons (31% of tissue-specific exons were predicted to have 50% or more disordered residues, compared to 25 and 16% of cassette and constitutive exons, respectively). The difference in the number of disordered exons was significant when tissue-specific exons were compared to both cassette and constitutive exons ($P<5.1x10^{-7}$ and $P<2.2x10^{-16}$, respectively, Chi-square test, where the value of $2.2x10^{-16}$ is the smallest P-value in R). Furthermore, to investigate whether protein disorder is in general a feature of genes that undergo tissue-specific splicing or if it is a specific characteristic of tissue-specific exons, I compared the fraction of disordered residues among the tissue-specific exons to the fraction of disordered residues in all other exons encoded by the representative transcripts with these exons. This showed that disordered residues are indeed characteristic for alternatively spliced tissue-specific exons (444 out of 1426 tissue-specific exons were encoding mostly disordered protein segment, compared to 3,543 out of 16,850 all other exons in these transcripts, $P<2.2x10^{-16}$, Chi-square test).

To ensure that observations about the prevalence of disordered residues in the tissue-specific exons are not biased by the applied disorder prediction method I used another method for identification of disordered regions. The VSL2B software is trained on datasets of disordered proteins and uses a linear support vector machine approach based on amino acid composition. Prediction of intrinsically disordered residues by this method confirmed that disordered residues are most common in the set of tissue-specific exons, followed by cassette exons. The fractions of exons with at least 50% predicted disordered residues were 53, 46 and 36% in the sets of tissue-specific, cassette and constitutive exons, respectively (Table 4.2). Hence, the observed enrichment of tissue-specific exons in disordered regions seems to be independent of the method for disorder prediction.

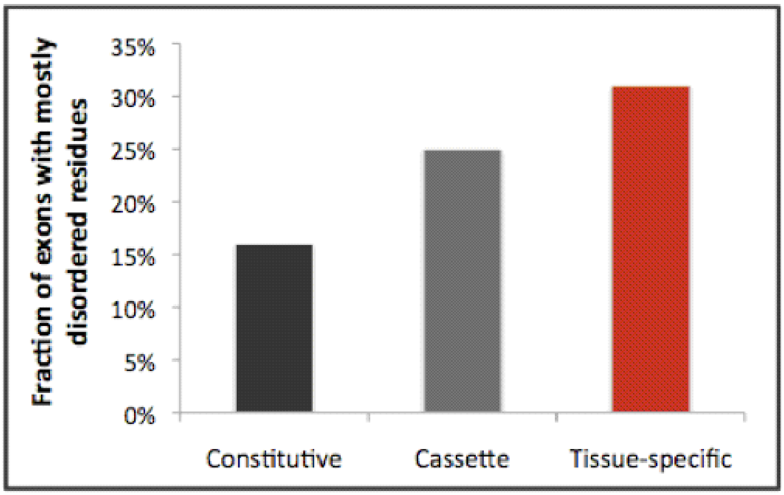

Figure 4.4: Protein regions encoded by tissue-specific exons are enriched in intrinsically disordered residues. The fraction of exons with at least 50% disordered residues in the three different sets of exons is shown. The number of exons with mostly disordered residues was significantly higher among tissue-specific exons when compared to cassette and constitutive exons ($P=5.1x10^{-7}$ and $P<2.2x10^{-16}$, respectively, Chi-square test, details in Appendix C.1). Disordered residues were predicted with the IUPred software.

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

Analysis	Set of exons	Fraction of disordered exons	P-value
VSL2	Tissue-specific	53%	/
	Cassette	46%	$P<5.4x10^{-8}$
	Constitutive	36%	$P<2.2x10^{-16}$

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

My hypothesis in this study is that disordered regions encoded by tissue-specific exons expose functional protein segments (Romero et al., 2006). Alternatively, these regions could act as fillers between functional structured domains (Tress et al., 2008; Tress et al., 2007). Functional disordered residues are frequently used in transient interactions in the cell, since their intrinsic flexibility allows them to be readily accessible to the proteins they interact with (Gsponer and Babu, 2009). I investigated here whether tissue-specific disordered residues indeed encode segments that could be used in protein interactions. Possible short protein binding sites and sites of post-translational modifications (PTMs) reflect disordered protein regions. Here, I analyzed whether there is evidence for a connection between tissue-specific disordered regions and protein binding sites. Firstly, I investigated whether unstructured segments contained peptide motifs that were likely to be bound by other proteins. For this, I used the ANCHOR software (Meszaros et al., 2009), which identifies disordered regions with a potential to bind protein domains on the hypothetical interaction partners. I found enrichment for the predicted functional peptide motifs in the tissue-specific exons compared to cassette and constitutive exons ($P<2.2x10^{-16}$

and $P<2.2x10^{-16}$, respectively, Chi-square test). Among the tissue-specific exons, 29% had a binding motif, compared to 18% of cassette exons and 18% of constitutive exons, see Figure 4.5a.

In addition, I investigated whether PTM sites were enriched in tissue-specific exons. For this, I looked at the annotated PTM sites in the Swiss-Prot portion of the UniProt database (Consortium, 2009). The analysis covered phosphorylation, methylation, acetylation and other PTM sites (Methods). This revealed that enrichment of PTM sites was indeed present in the set of tissue-specific exons. Tissue-specific exons had significantly more predicted PTM sites than cassette and constitutive exons ($P<9.9x10^{-12}$ and $P<3.2x10^{-7}$, respectively, Chi-square test). Among the tissue-specific exons from those transcripts that were successfully mapped to the UniProt isoforms, 13% had a PTM, compared to 7 and 8% of cassette and constitutive exons, respectively, see Figure 4.5b. PTM sites are frequently associated with unstructured regions (Holt et al., 2009; Iakoucheva et al., 2004) and in the set of tissue-specific exons, the majority (69%) of exons with at least one PTM site had a PTM in the predicted disordered region.

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

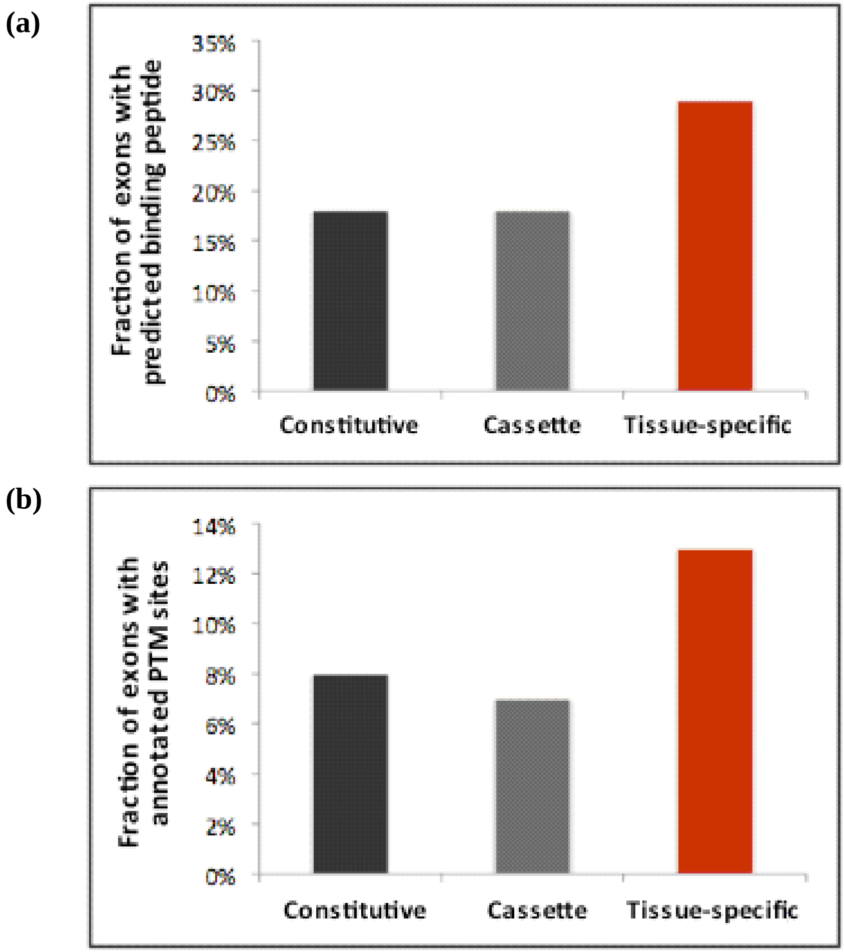

Figure 4.5: Tissue-specific exons encode protein segments enriched with predicted binding motifs and annotated PTM sites. (a) Fraction of exons with encoded binding motifs in the three different sets of exons. Binding motifs were predicted with the ANCHOR software. Tissue-specific exons were found to have a significantly higher fraction of predicted binding motifs than cassette and constitutive exons ($P<2.2x10^{-16}$ and $P<2.2x10^{-16}$, respectively, Chi-square test, details in Appendix C.1). (b) Fraction of exons with annotated PTM sites in the three different sets of exons. Tissue-specific exons were found to have a significantly higher fraction of PTM sites than cassette and constitutive exons ($P<=9.9x10^{-12}$ and $P<=3.2x10^{-7}$, respectively, Chi-square test, details in Appendix C.1). Positions of PTM sites in proteins were taken from the Swiss-Prot portion of the UniProt database.

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

Analysis	Set of exons	Fraction of exons with a phosphosite	P-value
Phosida phosphosites	Tissue-specific	2.3%	/
	Cassette	0.4%	$P<2.2x10^{-16}$
	Constitutive	0.5%	$P<2.2x10^{-16}$

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

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

I filtered out shorter exons from the sets of cassette and constitutive exons in order to compose tests sets with the average length of exons of 68 nucleotides. I compared the fractions of exons with predicted binding peptides and PTM sites in these two test sets with the one in the set of tissue-specific exons. I found that the set of tissue-specific exons still encoded a significantly higher fraction of PTM sites then the two test sets (Table 4.4). With regard to predicted binding peptides, I found that their fraction was significantly higher among the tissue-specific exons when compared to constitutive exons, but the difference was not that dramatic when compared to cassette exons (Table 4.4). Cassette exons have a higher fraction of disordered regions, so in that sense, it is not surprising that disordered binding motifs are more frequently predicted in that set than in the set of constitutive exons. However, overall, the analysis of subsets with longer cassette and constitutive exons confirmed that the observed enrichment of tissue-specific exons with functional sites is independent of the exon length.

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

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

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

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

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

Next, I looked at the conservation of predicted binding peptides only. Conservation of binding peptides was higher than the overall conservation of exons in all three sets, and it was the highest in the set of tissue-specific exons. Importantly, predicted binding residues were not only significantly more conserved in the tissue-specific exons when compared to cassette and constitutive exons ($P<2.2\text{x}10^{-16}$ and $P<2.2\text{x}10^{-16}$, respectively, Mann-Whitney test, Table 4.5), but were significantly more conserved then all other residues in the tissue-specific exons alone ($P=6.3\text{x}10^{-6}$, Mann-Whitney test, Table 4.6). Thus, even though the binding function of these residues is only predicted, it is likely that they play an important role in these proteins. The median value of conserved predicted binding peptides was 0.91 in tissue-specific exons, 0.86 in cassette exons and 0.86 in constitutive exons (Figure 4.6).

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

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

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

Set for analysis	Set of exons	N_{exons}	Median	P-value
Whole exons	Tissue-specific	1,404	0.89	N/A
	Cassette	44,750	0.86	$P<2.2\ x10^{-16}$
	Constitutive	143,811	0.87	$P<2.2\ x10^{-16}$
Disordered regions	Tissue-specific	883	0.90	N/A
	Cassette	24,120	0.83	$P<2.2\ x10^{-16}$
	Constitutive	68,719	0.84	$P<2.2\ x10^{-16}$
Binding peptides	Tissue-specific	630	0.91	N/A
	Cassette	13,600	0.86	$P<2.2\ x10^{-16}$
	Constitutive	37,708	0.86	$P<2.2\ x10^{-16}$

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

Set for analysis	Residues	N_{exons}	Median	P-value
Tissue-specific exons	Binding peptides	630	0.91	$P<26.3\ x10^{-6}$
	Other	1,363	0.89	

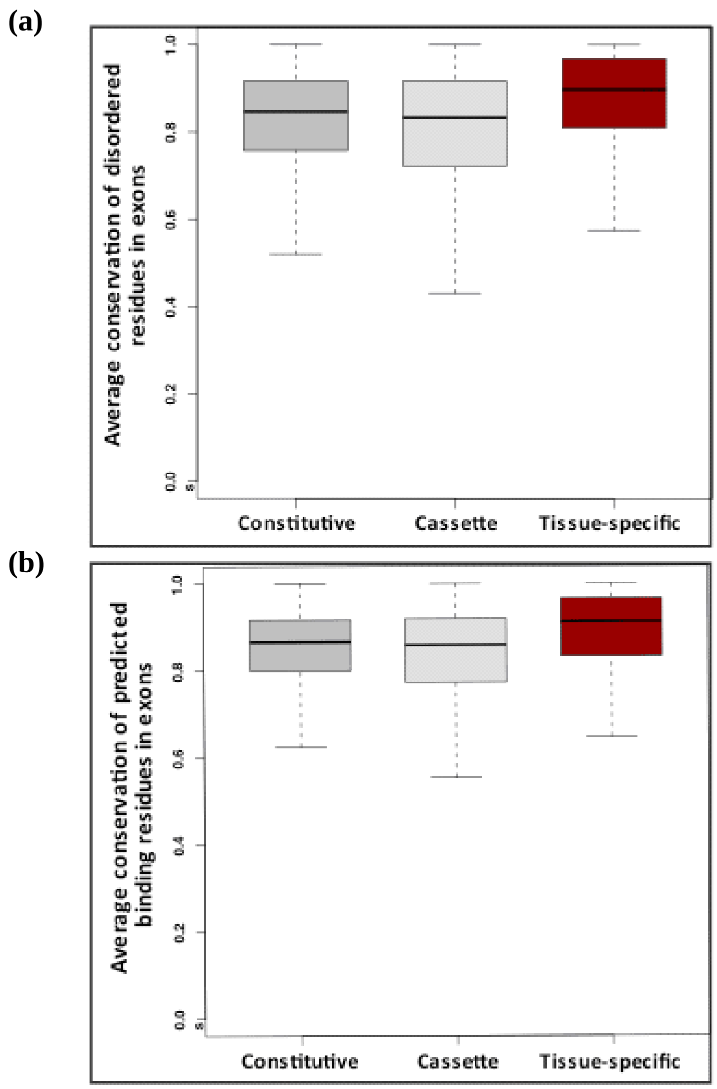

Figure 4.6: Residues in predicted disordered regions and peptide binding sites in the tissue-specific exons are highly conserved. (a) Conservation of predicted disordered residues in the three sets of exons, as calculated from residues aligned with mouse orthologus sequences. The median value for each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

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

If genes with tissue-specific isoforms tend to take part in different cellular pathways then mutations in these proteins are likely to have severe effects on the cellular and organism phenotype. I performed several analyses to see if this was the case. Firstly, I investigated whether genes from the MGI database (Bult et al., 2008), which are known to cause embryonic lethality in mice when mutated, were enriched with orthologues of human genes that have tissue-specific isoforms. I indeed found that genes with the tissue-specific isoforms were overrepresented among the genes involved in embryonic lethality ($P<1.2x10^{-8}$, Chi-square test, Table 4.7, Figure 4.7), which implied their potentially important role in the early stages of development.

Secondly, I investigated whether mutations in these genes could be related to cancer phenotype, since disruption of signalling pathways is a common initiator of the disease. Moreover, the study by Wang *et al*. that reported tissue-specific exons also included five different cancer cell lines, which increased the chances of detecting genes whose isoforms were potentially related to cancer. Indeed, I found that both Cancer Gene Census genes (Futreal et al., 2004) (genes that have been causally implicated in cancer) and genes from the COSMIC database (Forbes et al., 2008) (genes found to be somatically mutated in different cancer cells) were enriched with genes that have tissue-specific isoforms (P-values were $6.2x10^{-2}$ and $3.2x10^{-6}$ respectively, Chi-square test, Table 4.7, Figure 4.7). This suggested a possible connection between the genes with tissue-specific isoforms and cancer phenotype.

Finally, I investigated whether the genes with tissue-specific isoforms were enriched in any particular cellular pathway since this could possibly imply their influence on the phenotype. I found that these genes were enriched with genes that belong to the PDZ pathway (Table 4.8), a pathway in which disordered residues are known to play an important role. Apart from the significant overrepresentation of genes from PDZ pathway, this analysis revealed another important link; clustering of genes with similar function showed overrepresentation of genes from the MAPK pathway (Table 4.8). A possible connection with disordered residues here is suggested by the following example from the literature. The MAPK kinase MEK exists in two gene copies, *MEK1* and *MEK2*, which have essentially identical sequences but significantly different effects on the phenotype. I looked at the predicted disordered residues in these proteins and found that 54% of amino acids that differed between MEK1 and MEK2 were predicted to be disordered, compared to only 1% of the identical residues. Therefore, it is possible that in this known example from the MAPK pathway disorder functions as a mediator of protein interactions in a similar way in which I expect it acts in tissue-specific isoforms analysed here.

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

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

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

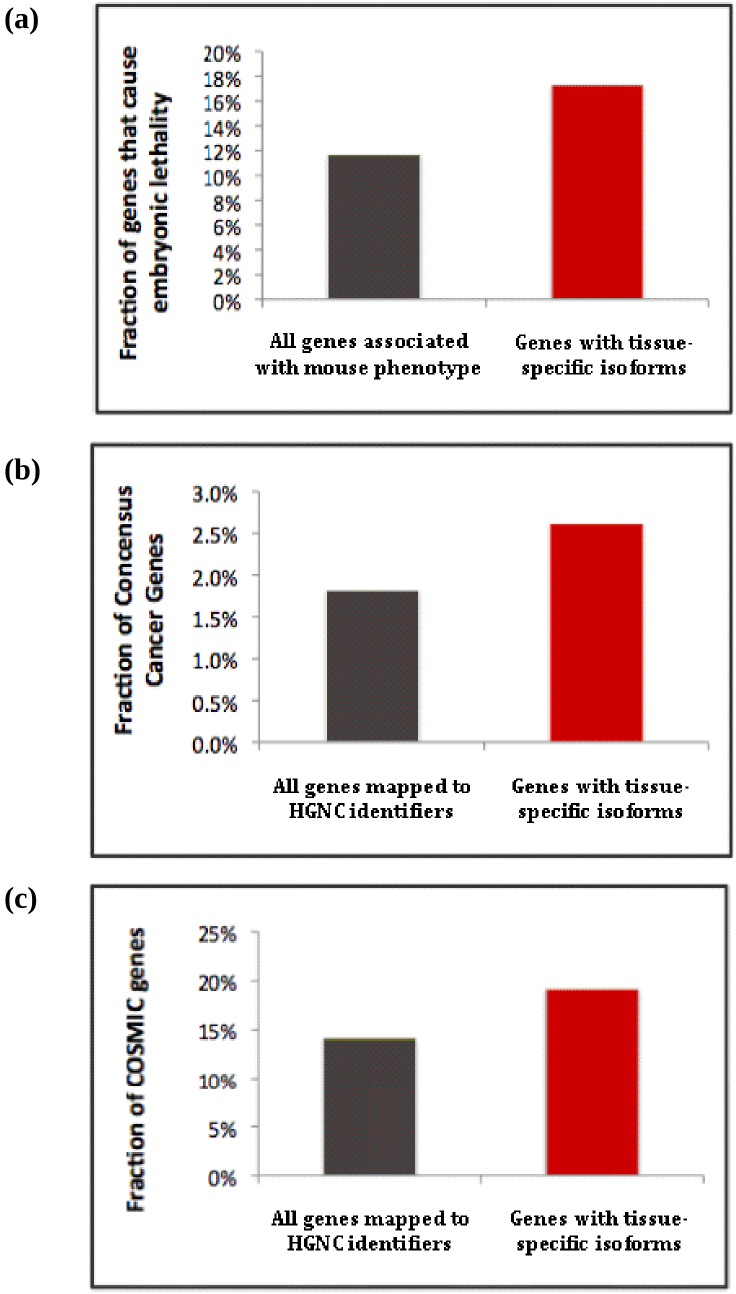

Figure 4.7: Fraction of genes with tissue-specific isoforms that are among the disease causing genes compared to background human genes. This is an illustration of data from Table 4.7. Background genes are always composed of all human genes with the identifiers in the corresponding databases. (a) Fraction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

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

Pathway	EASE P-value	Benjamini P-value
Enriched individual pathway:		
PDZ pathway: Synaptic Proteins at the Synaptic Junction	$2.3x10^{-5}$	$7x10^{-3}$
Enriched cluster of pathways with similar gene members:		
Mapk pathway: MAPKinase Signalling Pathway	0.06	0.90
P38 mapk pathway: p38 MAPK Signalling Pathway	0.26	0.98
Erk Pathway: Erk1/Erk2 MAPK Signalling pathway	0.35	0.99

4.3.7 Alternative isoforms of the gene p73

An example from the literature that illustrates the potential importance of alternative inclusion of exons that encode disordered protein segments is the one of the *p73* gene. Gene *p73* is a homologue of the *p53* gene and its main function is tumour suppression. However, this gene encodes a number of splice variants (Figure 4.8) which have been shown to be expressed in a tissue-specific manner (Ishimoto et al., 2002). These different splice isoforms all share the same central DNA binding region and differ in the alternative inclusion of N- and C-terminal exons (Bourdon, 2007). Functionally, the isoforms differ in their binding specificity, and the most striking of them is the ΔNp73 isoform which lacks the first three exons that encode the 'transactivating region' (Figure 4.8). Instead of acting as a tumour suppressor, the ΔNp73 isoform acts as an oncogene - possibly by competing with both p53 and other p73 isoforms for the DNA binding site (Ishimoto et al., 2002). When I predicted disordered regions (Dosztanyi et al., 2005) in the main protein isoform TP73001, which includes also the terminal exons, I observed that the encoded protein had several disordered segments and most importantly, that the N-terminal region encoded by the first three exons is predominantly disordered (Figure 4.8). Additionally, this region also contained two predicted binding peptides (not shown), as predicted by ANCHOR (Meszaros et al., 2009). Similarly to the p73 protein, it has been reported previously that the N-terminal region of the human p53 tumour suppressor protein contained large disordered segments (Bell et al., 2002; Dawson et al., 2003). The N-terminal part of the p53 protein has an important regulatory role (Chumakov, 2007), and so far, three different protein partners have been shown to bind to this region – binding peptides for these proteins were successfully predicted with ANCHOR (Meszaros et al., 2009). The example of the p73 gene clearly illustrates that the alternative inclusion of disordered protein segments can dramatically affect the function of a protein.

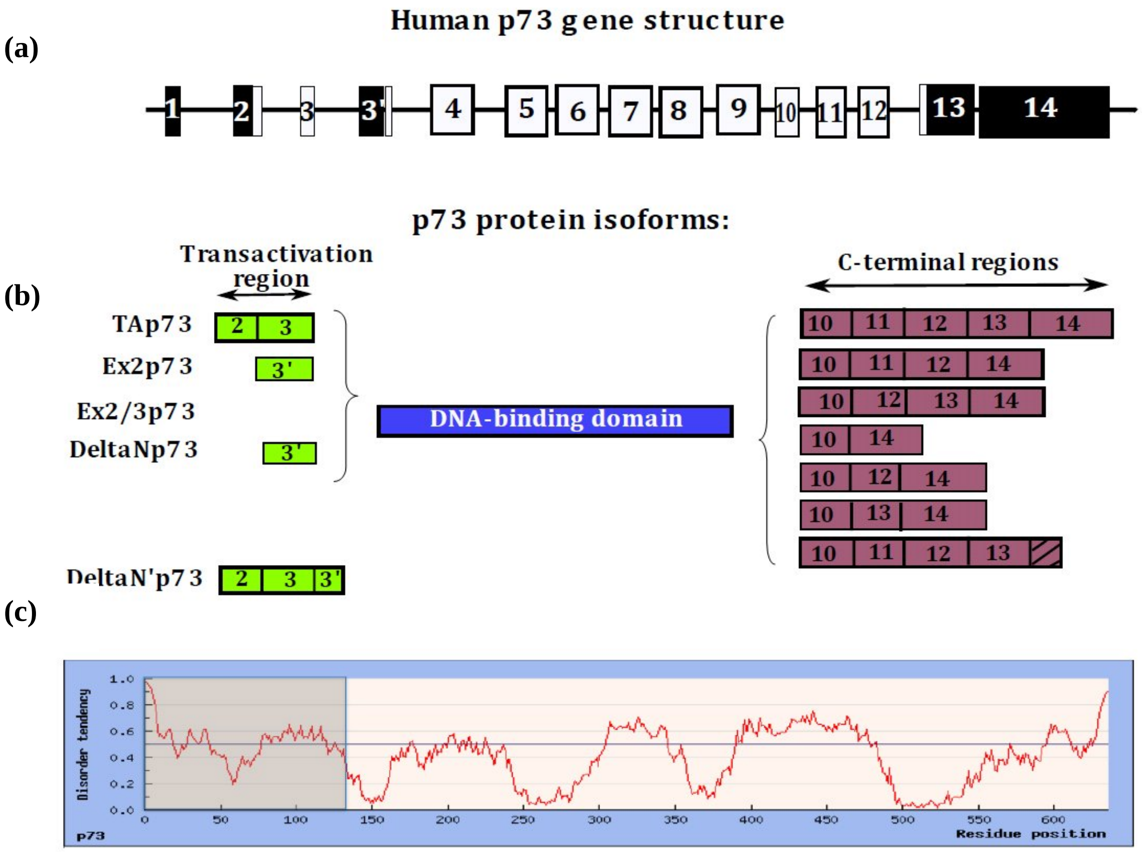

Figure 4.8: *p73* gene and its isoforms. Figures a and b are adapted from (Bourdon, 2007), and show different splicing events that occur in the *p73* gene. (a) The intron-exon structure of the gene is shown. Black boxes indicate 5' and 3' untranslated exon regions, and white boxes exon regions that encode protein sequences. There are two alternative transcription start sites: before the exon 1 and before the exon 3'. (b) Protein segments encoded by different exons are shown. All splice isoforms apart from ΔN'p73 have the central DNA-binding domain, but differ in the segments encoded by the N- and C-terminal exons. The exon numbering in section (a) is transferred to section (b) of the figure. (c) Disordered regions (threshold is at 0.5 disorder tendency) in the longest Tap73 (*TP73-001*) isoform, as predicted with the IUPred software. This isoform includes the first three N-terminal exons that are missing in the ΔNp73 isoform and these are indicated with a grey square in the disorder prediction graph. The greatest part of the protein segment encoded by these first three exons is predicted to be disordered.

4.3.8 Tissue-specific splicing and protein domains

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

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

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

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

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

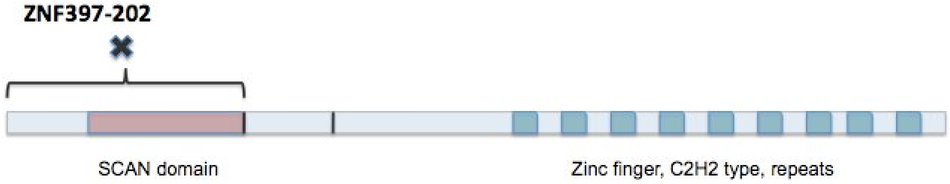

Figure 4.9: Example of a tissue-specific exon exclusion that removes the whole domain from a protein product. The *ZNF397-202* isoform of Zinc finger protein 397 (ENST00000330501) encodes several DNA binding Zinc finger repeats and the protein interaction SCAN domain. Tissue-specific alternative splicing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

4.4 Discussion

4.4.1 Evolution and function of alternative splicing

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

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

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

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

This study focused on alternatively spliced exons that encode functional residues which determine protein-protein interactions. However, alternative inclusion of other, even short, protein segments can have dramatic consequences for the overall protein function. A good illustration for this is the Piccollo protein (Garcia et al., 2004). This gene produces two protein isoforms that differ in nine residues. As a result of this, the shorter isoform has a stronger binding affinity for Ca^{2+}, but is also incapable of undergoing Ca^{2+}- dependent dimerization that normally occurs in a longer isoform. The structural study of this protein showed that this was a consequence of a large structural change induced by the omitted short motif. Apart from causing a drastic change in protein structure, alternative splicing can also affect the connector region between the globular domains of a protein and in that way influences their orientation and recruitment of their binding partners. Additionally, splicing can also affect regions that determine ligand binding, which was not covered in this study. Hence, design of this study covers only a fraction of alternative splice events that can have important consequences for the overall protein function.

4.4.2 Unstructured functional residues direct isoform-specific networks

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

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

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

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

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

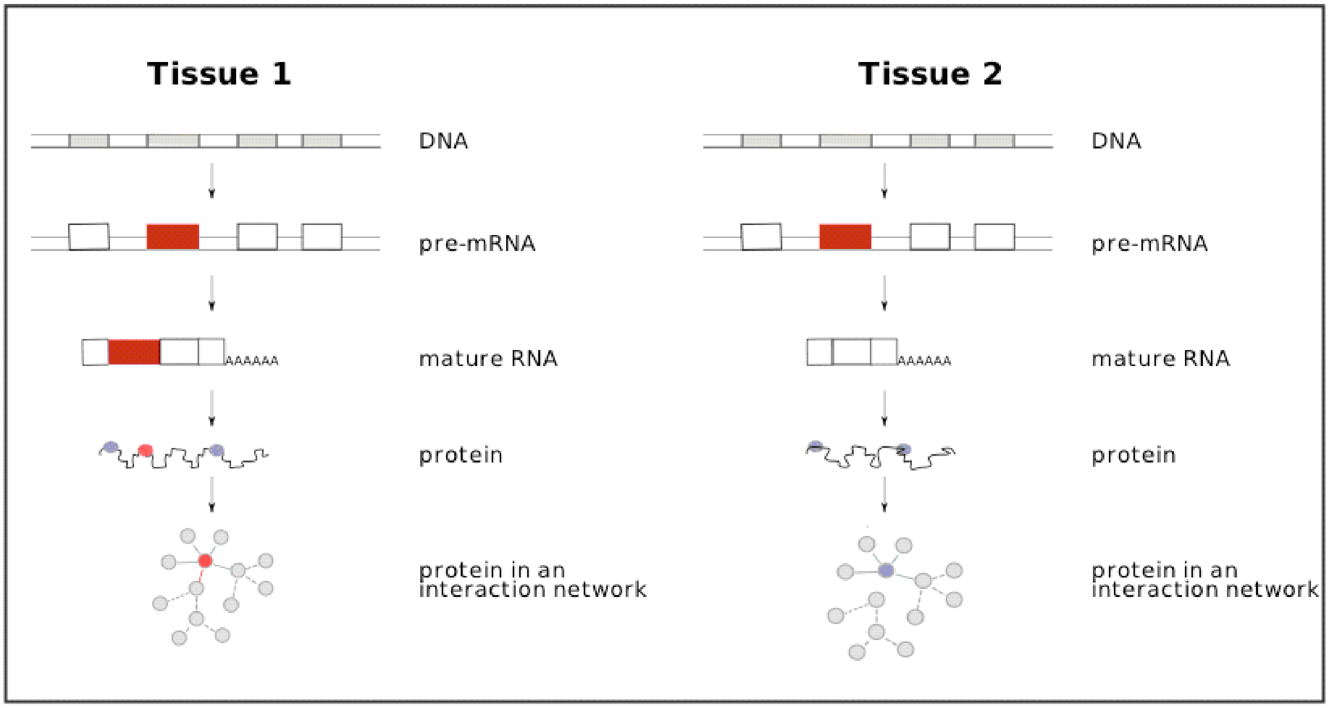

Figure 4.11: Illustration of the predicted effect of tissue-specific alternative splicing of functional disordered residues. Tissue-specific splicing and differential inclusion of exons frequently results in differential presence of a protein segment with specific binding motifs. Binding motifs are shown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

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

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

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

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

Genes with tissue-specific isoforms are enriched in genes from the PDZ pathway but are also reported to include genes from several pathways related to MAPK signalling (Table 4.8). As discussed in the introduction, this central signalling pathway can activate numerous cellular processes and represents a good hypothetical target for modulation of protein function through alternative inclusion of disordered binding residues. However, the role of functional disordered residues has not been connected with this pathway so far. Nonetheless, the example of the MEK kinase shows disorder could be utilized in this pathway to direct specific signalling. The MEK kinase exists in two gene copies: *MEK1* and *MEK2*. Sequences of their protein products are highly similar and their kinase domains essentially identical; they were initially even considered to be functionally redundant (Shaul and Seger, 2007). However, the proteins do differ in their N-termini and in the proline-rich inserts (residues phosphorylated by MAPK kinase kinases). As a result, each protein forms signalling complexes with different protein partners (Shaul and Seger, 2007). This has such strong implications that knockout of MEK1 causes an embryonic lethality in MEK1$^{-/-}$ mice whereas MEK2$^{-/-}$ mice are viable and fertile (Shaul and Seger, 2007). The analysis of the MEK1 and 2 protein sequences showed that their N-terminal regions are indeed unstructured (section 4.3.6).

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

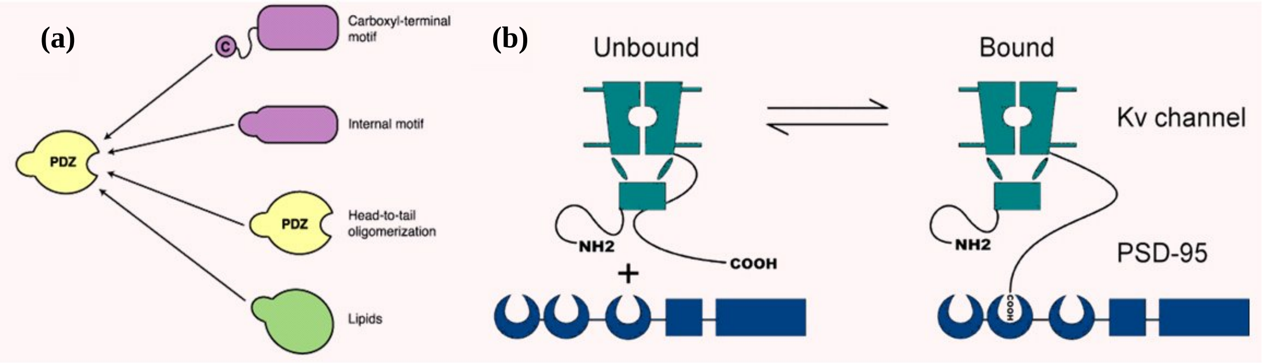

Figure 4.12: PDZ domain proteins play an important role in the targeting of proteins to specific membrane compartments and their assembly into supramolecular complexes. a) PDZ domains participate in at least four different classes of interaction: recognition of C-terminal motifs in peptides, recognition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

4.4.4 Genes with tissue-specific isoforms and disease development

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

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

By inclusion of disordered regions, functional capability of a single protein can expand depending on the context, space and time. When this process is related to disease development, it is an attractive target for drug application - especially if a drug, such as for example an antibody, can be made specific for one isoform and not interfere with the function of other isoforms. However, in order to be able to interfere with this process, it is necessary first to understand it. More comprehensive studies of splicing and genomic architecture in an increasing number of species will surely play an important role in addressing this problem.

4.5 Bibliography

Bell, S., Klein, C., Muller, L., Hansen, S., and Buchner, J. (2002). p53 contains large unstructured regions in its native state. J Mol Biol *322*, 917-927.

Bochkareva, E., Kaustov, L., Ayed, A., Yi, G.S., Lu, Y., Pineda-Lucena, A., Liao, J.C., Okorokov, A.L., Milner, J., Arrowsmith, C.H., *et al.* (2005). Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. Proc Natl Acad Sci U S A *102*, 15412-15417.

Bourdon, J.C. (2007). p53 and its isoforms in cancer. Br J Cancer *97*, 277-282.

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., and Dunker, A.K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol *55*, 104-110.

Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Blake, J.A. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res *36*, D724-728.

Chen, J.W., Romero, P., Uversky, V.N., and Dunker, A.K. (2006). Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. J Proteome Res *5*, 888-898.

Chumakov, P.M. (2007). Versatile functions of p53 protein in multicellular organisms. Biochemistry (Mosc) *72*, 1399-1421.

Consortium, U. (2009). The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res *37*, D169-174.

Dawson, R., Muller, L., Dehner, A., Klein, C., Kessler, H., and Buchner, J. (2003). The N-terminal domain of p53 is natively unfolded. J Mol Biol *332*, 1131-1141.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*, P3.

Di Lello, P., Jenkins, L.M., Jones, T.N., Nguyen, B.D., Hara, T., Yamaguchi, H., Dikeakos, J.D., Appella, E., Legault, P., and Omichinski, J.G. (2006). Structure of the Tfb1/p53 complex: Insights into the interaction between the p62/Tfb1 subunit of TFIIH and the activation domain of p53. Mol Cell *22*, 731-740.

Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I., and Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. J Proteome Res *5*, 2985-2995.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J *272*, 5129-5148.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol *6*, 197-208.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., *et al.* (2008). The Pfam protein families database. Nucleic Acids Res *36*, D281-288.

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., and Stratton, M.R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet *Chapter 10*, Unit 10 11.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. Bioinformatics *23*, 950-956.

Garcia, J., Gerber, S.H., Sugita, S., Sudhof, T.C., and Rizo, J. (2004). A conformational switch in the Piccolo C2A domain regulated by alternative splicing. Nat Struct Mol Biol *11*, 45-53.

Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol *8*, R250.

Gsponer, J., and Babu, M.M. (2009). The rules of disorder or why disorder rules. Prog Biophys Mol Biol *99*, 94-103.

Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science *322*, 1365-1368.

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., and Iakoucheva, L.M. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol *2*, e100.

Holt, L.J., Tuch, B.B., Villen, J., Johnson, A.D., Gygi, S.P., and Morgan, D.O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. Science *325*, 1682-1686.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol *4*, R70.

Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., *et al.* (2009). Ensembl 2009. Nucleic Acids Res *37*, D690-697.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., and Dunker, A.K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol *323*, 573-584.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res *32*, 1037-1049.

Ishimoto, O., Kawahara, C., Enjo, K., Obinata, M., Nukiwa, T., and Ikawa, S. (2002). Possible oncogenic potential of DeltaNp73: a newly identified isoform of human p73. Cancer Res *62*, 636-641.

Jin, P., Fu, G.K., Wilson, A.D., Yang, J., Chien, D., Hawkins, P.R., Au-Young, J., and Stuve, L.L. (2004). PCR isolation and cloning of novel splice variant mRNAs from known drug target genes. Genomics *83*, 566-571.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet *11*, 345-355.

Kim, E., and Sheng, M. (2004). PDZ domain proteins of synapses. Nat Rev Neurosci *5*, 771-781.

Kondrashov, F.A., and Koonin, E.V. (2003). Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends Genet *19*, 115-119.

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., *et al.* (2009).

ASTD: The Alternative Splicing and Transcript Diversity database. Genomics *93*, 213-220.

Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing. Trends Genet *19*, 124-128.

Kussie, P.H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A.J., and Pavletich, N.P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science *274*, 948-953.

Lareau, L.F., Green, R.E., Bhatnagar, R.S., and Brenner, S.E. (2004). The evolving roles of alternative splicing. Curr Opin Struct Biol *14*, 273-282.

Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., Leibman-Barak, S., Pupko, T., and Ast, G. (2007). The "alternative" choice of constitutive exons throughout evolution. PLoS Genet *3*, e203.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet *24*, 416-425.

Magidovich, E., Orr, I., Fass, D., Abdu, U., and Yifrach, O. (2007). Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K+ channel modulates its interaction with scaffold proteins. Proc Natl Acad Sci U S A *104*, 13022-13027.

Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS Comput Biol *5*, e1000376.

Modrek, B., and Lee, C.J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet *34*, 177-180.

Nourry, C., Grant, S.G., and Borg, J.P. (2003). PDZ domain proteins: plug and play! Sci STKE *2003*, RE7.

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell *127*, 635-648.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics *7*, 208.

Ranganathan, R., and Ross, E.M. (1997). PDZ domain proteins: scaffolds for signaling complexes. Curr Biol *7*, R770-773.

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., and Lee, C. (2004). Assessing the impact of alternative splicing on domain interactions in the human proteome. J Proteome Res *3*, 76-83.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., *et al.* (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A *103*, 8390-8395.

Shaul, Y.D., Gibor, G., Plotnikov, A., and Seger, R. (2009). Specific phosphorylation and activation of ERK1c by MEK1b: a unique route in the ERK cascade. Genes Dev *23*, 1779-1790.

Shaul, Y.D., and Seger, R. (2007). The MEK/ERK cascade: from signaling specificity to diverse functions. Biochim Biophys Acta *1773*, 1213-1226.

Shimizu, K., and Toh, H. (2009). Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. J Mol Biol *392*, 1253-1265.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005). Function of alternative splicing. Gene *344*, 1-20.

Taylor, J., Schenck, I., Blankenberg, D., and Nekrutenko, A. (2007). Using galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics *Chapter 10*, Unit 10 15.

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett *579*, 3346-3354.

Tompa, P., Szasz, C., and Buday, L. (2005). Structural disorder throws new light on moonlighting. Trends Biochem Sci *30*, 484-489.

Tress, M.L., Bodenmiller, B., Aebersold, R., and Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol *9*, R162.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Olason, P.I., Albrecht, M., Hegyi, H., Giorgetti, A., *et al.* (2007). The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci U S A *104*, 5495-5500.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res *64*, 7647-7654.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, K., Geren, L., Zhen, Y., Ma, L., Ferguson-Miller, S., Durham, B., and Millett, F. (2002). Mutants of the CuA site in cytochrome c oxidase of Rhodobacter sphaeroides: II. Rapid kinetic analysis of electron transfer. Biochemistry *41*, 2298-2304.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol *293*, 321-331.

Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol *11*, R53.

Xing, Y., and Lee, C.J. (2005). Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genet *1*, e34.

Yura, K., Shionyu, M., Hagino, K., Hijikata, A., Hirashima, Y., Nakahara, T., Eguchi, T., Shinoda, K., Yamaguchi, A., Takahashi, K., *et al.* (2006). Alternative splicing in human transcriptome: functional and structural influence on proteins. Gene *380*, 63-71.