Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

The impact of splicing on protein domain architecture

Sara Light[1,2,3] and Arne Elofsson[1,2,4]

Many proteins are composed of protein domains, functional units of common descent. Multidomain forms are common in all eukaryotes making up more than half of the proteome and the evolution of novel domain architecture has been accelerated in metazoans. It is also becoming increasingly clear that alternative splicing is prevalent among vertebrates. Given that protein domains are defined as structurally, functionally and evolutionarily distinct units, one may speculate that some alternative splicing events may lead to clean excisions of protein domains, thus generating a number of different domain architectures from one gene template. However, recent findings indicate that smaller alternative splicing events, in particular in disordered regions, might be more prominent than domain architectural changes.The problem of identifying protein isoforms is, however, still not resolved. Clearly, many splice forms identified through detection of mRNA sequences appear to produce 'nonfunctional' proteins, such as proteins with missing internal secondary structure elements. Here, we review the state of the art methods for identification of functional isoforms and present a summary of what is known, thus far, about alternative splicing with regard to protein domain architectures.

Addresses
[1] Science for Life Laboratory, Stockholm University, Box 1031 SE-171 21 Solna, Sweden
[2] Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
[3] Bioinformatics Infrastructure for Life Sciences (BILS), Sweden
[4] Swedish e-Science Research Center (SeRC), Sweden

Corresponding author: Elofsson, Arne (arne@bioinfo.se, arne.elofsson@gmail.com)

Current Opinion in Structural Biology 2013, **23**:451–458

This review comes from a themed issue on **Sequences and topology**

Edited by **Julian Gough** and **A Keith Dunker**

For a complete overview see the Issue and the Editorial

Available online 4th April 2013

http://dx.doi.org/10.1016/j.sbi.2013.02.013.

Introduction — domain architectures and splicing

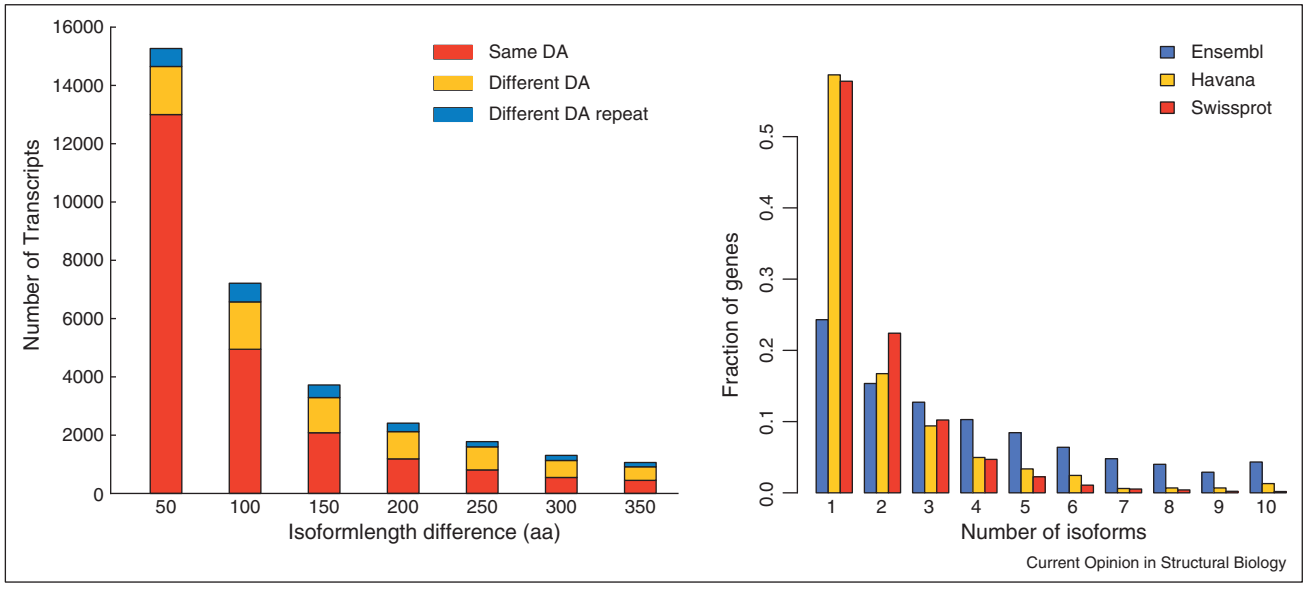
Protein domains are structural, functional and evolutionary building blocks that, within one protein, can form various architectures that may be composed of one or several domains [1]. Domains can often be defined either from a sequence similarity viewpoint as in the Pfam database [2], from an evolutionary perspective as in SCOP [3] or from a structural perspective as in CATH [4]. In many cases these definitions overlap [5].

Early in the genomic era studies showed that multidomain proteins are much more common in eukaryotes than prokaryotes [6] and that about two-thirds of eukaryotic proteins consist of two or more domains [7]. Novel multidomain architectures have been created primarily by single domain additions at the N-terminus or C-terminus of proteins [8] and the increase in novel architectures in Metazoa [9] can at least partly be explained by a set of metazoan specific exon-bordering domains [10]. However, these observations might, in part, suffer from errors generated by gene prediction [11] and also from protein relationships by epaktology, that is proteins only related through shared domains [12].

Metazoan genes are much more complex than the genes of simpler organisms and are therefore quite difficult to annotate correctly. Fundamental to our understanding of splicing is our understanding of introns and exons. Introns are common in the genomes of almost all higher organisms, while virtually nonexistent in prokaryotes and quite rare in fungi [13]. One question that has been debated for years is whether introns arose before the split between prokaryotes and eukaryotes or after [14]. In a recent paper by Rogozin *et al.* [15] the authors argue that many introns are shared between distant eukaryotes, and therefore, most likely, were present in the earliest eukaryotes. Subsequently, much of the variation in intron content seen between different organisms is primarily due to a loss of introns. However, the authors also observed that there is a rapid, albeit temporary, increase in introns around the origin of Metazoa and, further, suggest that alternative splicing is predominantly due to *splicing errors* rather than the result of a deterministic process. Regardless, alternative splicing provides a major contribution to the biological complexity of multicellular eukaryotes.

Splicing has long been recognized as a likely source of added phenotypic complexity [16–19]. Splicing patterns vary from cell to cell in complex organisms and many examples of functionally important splice forms have been reported, see for instance a recent review by Kelemen *et al.* [20]. However, although more than 22,000 articles in PubMed contain the phrase 'alternative splicing', the function of the vast majority of splice forms is not known [21]. Most of the human protein coding genes can produce alternatively spliced mRNAs [22,23,24•], and for human genes the number of transcripts is often larger than three [25••] (Figure 1). However, the abundance of

Figure 1



The left panel shows the fraction of proteins where the domain architecture is altered as a result of splicing. The plot shown is based on Swissprot transcripts [33]. Swissprot is the manually curated portion of the UniprotKB database. The same calculation performed on Vega/Havana [30], Ensembl and Uniprot show the same general trends (data not shown). The right panel shows the number of isoforms for three databases; Ensembl, Vega/Havana and Swissprot.

transcripts may not be translated directly to an abundance of protein functions [25••].

Several mechanisms producing alternative protein forms are briefly described in Figure 2. Splicing is mediated by a large molecular machinery, the spliceosome, that recognizes the exons by three major sequence elements: the 5′ splice site, the 3′ splice site and a branch point [20]. The 5′ splice site consist of an AG base pair and is often preceded by a non-AG region, which tends to be more extended in alternatively spliced exons than in other exons. Such signals could potentially be used to identify conserved splice signals between organisms. Initially, splice junction microarrays were primarily used to quantify splice variants, but given the rapid progress in sequence technology, RNAseq is increasingly the predominant method [26]. It has been shown that RNAseq generates identifiable gene models for a larger set of the genes than array platforms [27]. Additionally, it seems likely that the progress in the field of proteomics will shed light on the validity and biological functionality of alternatively spliced transcripts [28].

Alternative splicing in the human proteome

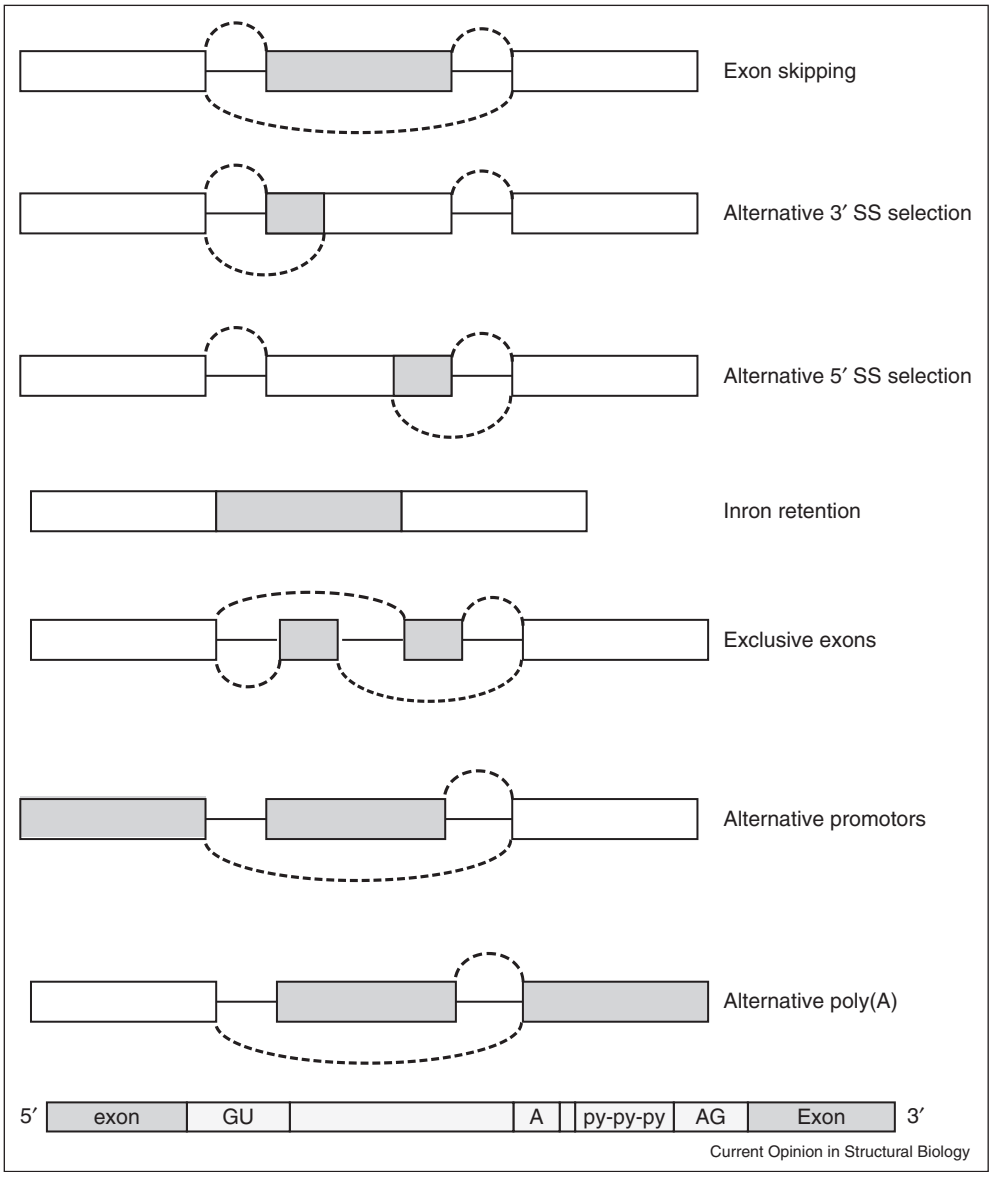
In the early days of genomics, many different dedicated alternative splicing databases were produced. However, to the best of our knowledge hardly any of these have been consistently updated during the last few years, so today the best resources for studying alternative splicing are the more general databases: firstly, Ensembl [29] — a database that contains eukaryotic genomes; secondly, Vega/Havana [30,31•] — a resource for vertebrate genome annotation; thirdly, Unigene [32] — a transcriptome database; fourthly, Uniprot [33] — the comprehensive protein database; and finally, Gencode [34] — the encyclopedia of genes and gene variants. Among the specialized databases of alternative splicing, few have stood the test of time, but there are two promising resources that are quite recent. First, ASPicDB [35••] — a database that provides access to reviewed annotations of alternative splicing for human genes and, second, APPRIS, a database that contains annotations of human isoforms [36•].

Ten years ago Kriventseva *et al.* [37] reviewed the state of splicing with respect to domain borders in Swissprot. In today's perspective this study is quite small only including 4,804 splicing variants of 1,780 proteins. In comparison, today (December 2012) Swissprot contains almost 15 000 spliced human proteins in more than 37,000 splice forms, see Figure 1. About 55% of the splice variants include a missing region in one variant and in the remaining 45% one region has been replaced. The missing regions are significantly longer (average 234 residues) than the replaced regions (32 residues). If we turn to Ensembl or Unigene the number of splice forms is considerably larger.

Identification of functional isoforms

It came as a surprise for many when, in 2007, Tress and co-authors [25••] first showed that alternative splicing is

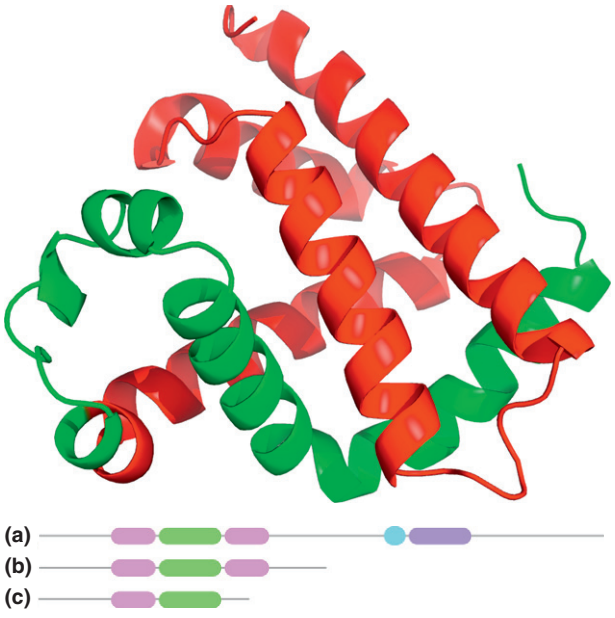
Figure 2



In the top a schematic illustration of different alternative splicing mechanisms are shown. The gray protein coding regions are excluded/included in different transcripts. This figure is inspired by Keren *et al.* [66]. The bottom figure shows the most important sequence patterns related to a splicing. This image is modified from the Wikipedia article on RNA splicing. The splicing start with an AG site and is preceded with a non-AG (pyrimidine rich) region preceded by the branch point that includes an Adenosine residue. The 5′ end of the intro contains an almost invariant GU sequence. In both images exons are marked in gray and intron in white.

even more common than previously thought. Further, the results indicated that for many of the alternative protein products, there is strong evidence suggesting that they encode nonfunctional proteins. Perhaps most strikingly, the authors suggested that it is unlikely that the 'spectrum of conventional enzymatic or structural functions can be substantially extended through alternative splicing'. The analysis was partly performed by homology modeling of alternative splicing products resulting in proteins that lacked central parts of their structure. Indeed, the vast majority of splice forms may occur due to noise in the splicing machinery [38]. Later, Tress *et al.* used an approach combining five different methods [39]: conservation of exonic structure; nonneutral evolution; protein structure mapping; functional residue conservation using firestar [40] and SQUARE [41] and vertebrate alignments. By utilizing these five tools they were able to identify the principal isoform for 83% of the proteins.

Figure 3



The top panel shows an example of splicing in myoglobin (ENSG00000198125) visualized using the MAISTAS tool [67•]. Here, only the two highly reliable, according to Ensembl, transcripts (ENSP00000380489 and ENSP00000386060) are visualized. In ENSP00000386060 the green part is missing and this isoform is noted by MAISTAS to have an exposed hydrophobic surface larger than expected and less optimal contacts, that is this protein isoform is unlikely to be folded. The bottom panel shows a schematic illustration of EGFR, epidermal growth factor receptor. Three reviewed isoforms that are both verified Havana transcripts and predicted. The domains include receptor L (pink), furin-like (green), transmembrane region (blue) and protein tyrosine kinase domain (purple).

A follow-up to the Tress study was performed in 2011 by Hegyi *et al.* [42•]. They used a novel approach — 'Domain Integrity Verification of Alternative Splicing' (DIVAS) — to identify 'functional' splice products. This strategy had earlier proved very useful for identifying mis-assigned proteins [43]. They noted that only 14 out of 4000 human proteins in PDB were associated with two (or more) isoforms while 95% of multiexonic human protein-coding genes undergo alternative splicing. Further, none of the splice forms differed by more than five amino acids, that is could not possibly contain an entire domain. They went on to analyze 505 human isoforms from Swissprot and found 'that strict rules govern the selection of alternative splice variants aimed to preserve the integrity of globular domains: alternative splice sites firstly, tend to avoid globular domains; secondly, affect them only marginally; thirdly, tend to coincide with a location where the exposed hydrophobic surface is minimal; or finally, the protein is disordered.' Here it should be noted that the selected 505 splice variants had literature evidence supporting their existence at the protein level. In the entire Swissprot set (that have mRNA evidence from more than one study) *22%* of the splice events affect a domain boundary (compared to 35% at random), a number that was merely 9% in the validated set.

A somewhat different picture emerges when alternatively spliced products are studied by mass spectrometry, a method that allows experimental verification of the presence of a protein in the cell. Here, Tress *et al.* showed that, in *Drosophila*, many stable alternatively spliced isoforms exist [28]. This was also confirmed in higher organisms [44,45••]. Another explanation for this observation is that proteins may be much more tolerant to structural deletions, insertions and replacements than previously thought [46] or the fact that many of the alternatively spliced transcripts code for protein regions that appear to be intrinsically disordered [47]. It has been proposed that an important function for alternatively spliced isoforms is to remodel the protein–protein interaction network [48], often mediated through intrinsically disordered regions.

From a recent study by Mudge *et al.* [49•], based on 309 protein coding genes from mouse and human with respect to splicing, utilizing transcriptomic and RNAseq data, it is clear that even splice forms associated with nonsense mediated decay (NMD), a regulatory process by which nonfunctional transcripts are degraded, can be evolutionarily conserved and, that is have a functional role, possible for regulating expressions levels [50].

With these studies in mind, it would seem that a majority of the splice products are unlikely to produce functional proteins. It has been assumed that these transcripts are instead targeted for nonsense mediated decay (NMD); however, to the best of our knowledge this has not been explicitly proven, but it is quite clear that these transcripts are nonfunctional. Obviously this causes many problems when analyzing alternative splicing and differences in domain architectures since it becomes crucial to identify the transcripts that are associated with protein isoforms.

How does alternative splicing affect the protein domain architecture?

After noting that, according to current consensus, only a small fraction of all alternatively spliced products result in functional proteins, it is obvious that it is crucial to correctly select the biologically relevant isoforms before performing an analysis of different splicing forms. Several different methods to limit the datasets have been explored. One approach is to use only conserved splice forms between, for instance, mouse and human.

In one of the first large scale studies of domains and alternative splicing Liu and Altman [51•] identified 24 domains that were significantly more common in proteins undergoing alternative splicing than in other human

proteins. The most over-represented domain was the repeating cadherin domain. Over-represented domains are predominantly involved in the processes of cell communication, signaling, development and apoptosis, both with regard to domains present in proteins undergoing alternative splicing and when it comes to 'spliced out domains'.

At roughly the same time Kriventseva *et al.* [37] showed that there was a selective pressure that serves to keep domain borders intact. However, still only 21% of the spliced regions overlapped with a domain border. These observations have been confirmed in later studies. They also noted that alternative splicing occurring inside protein domains preferentially targets functional amino acids and that entire domains are removed more frequently than expected by chance. Finally, as noted in the studies by Tress [25••], 60% of the alternative protein isoforms that they were able to model lacked long parts of a domain.

In 2004 Taneri *et al.* [52] studied alternative splicing on transcription factors in mouse. They found that in these proteins it is mainly DNA binding domains that are added or deleted in different isoforms, providing tissue specific variants.

In an attempt to predict functional isoforms Leoni *et al.* showed that the most effective strategy for correctly identifying translated products relies on the conservation of active sites [44]. However, this can only be applied to a small set of isoforms. A better coverage can be achieved by analyzing the presence of nontruncated functional domains, thus showing the importance of domains when studying alternative splicing.

In addition to the observation that some domains are more common in spliced proteins than others it has recently been observed that intrinsic protein disorder is common in spliced proteins [28,53]. Considering the prominence of disordered proteins among the hubs in the protein–protein interaction network [54] and the central role of disorder in signaling [55], it is possible that the functional reason for alternative splicing of intrinsically disordered regions is to rewire interaction networks [56].

We checked Swissprot and found that around 36% of the splice forms affect the domain architecture of the proteins, see Figure 1, and the corresponding number for Havana (Human And Vertebrate Analysis and Annotation) transcripts is 43%. However, although these transcripts are reviewed and considered reliable it is still, as mentioned above, unclear what fraction of these splice forms produce functional proteins.

Splicing and domain architecture for functional variation

There are some well studied examples where alternative splicing affects domain structure and clearly yields a domain architectural and/or phenotypic effect. Some of the best established examples of isoforms with domain architectural changes are associated with cancer such as for instance the epidermal growth factor receptor (EGFR), a transmembrane protein that belongs to the protein kinase family (Figure 3). This protein is, in various isoforms, overexpressed in many cancers [57]. The longest isoform contains, aside from a transmembrane region, four protein domains; two copies of the Receptor L domain and one copy each of the furin-like domain and a protein tyrosine kinase domain. There are three revised isoforms of this gene that contain different numbers of domains. Another example is collagen alpha-3 (VI), a protein of the extracellular matrix [58]. The main difference between the short isoforms of this protein and the longer one is that the latter contains a von Willebrand factor domain along with seven predicted phosphorylation sites [59]. This protein is prevalent in connective tissue and the longer isoform is nearly absent from normal tissue but is quite abundant in cancer samples.

Further, some protein domains that are associated with repeat proteins are common among alternative exons

Table 1

The ten most frequent domains that are associated with domain architecture (DA) differences between splice forms. The calculation is based on the Havana set of human isoforms from the Ensembl database, downloaded in December 2012. The numbers indicate the number of times the domain has been found to differ between splice forms

DA difference	Pfam ID	Description
170	CL0023	P-loop contain nucleoside triphosphate hydrolase superfamily
168	CL0159	Ig-like fold superfamily (E-set)
158	CL0011	Immunoglobulin superfamily
154	CL0020	Tetratrico peptide repeat superfamily
132	CL0126	Peptidase clan MA
124	CL0361	Classical C2H2 and C2HC zinc fingers
124	CL0123	Helix-turn-helix clan
96	PF08172.7	CASP C terminal
96	PF02376.10	CUT domain
90	PF00681.15	Plectin repeat

[51•], see Table 1, as for example in fibroblast growth receptor I where a immunoglobulin domain is alternatively spliced, thus affecting cellular proliferation [60].

Aside from generating isoforms with different domain architectures, protein domains themselves may also be modified through alternative splicing. This holds true for the Piccolo protein [61] — a protein, that is implicated in organizing neuronal zones — where a nine residue insert due to alternative splicing, occurring in the C_2A domain, changes the structural fold and leads to a markedly reduced affinity for calcium. Further, Weatheritt *et al.* recently showed an enrichment of short linear motifs among alternative exons, that lead to protein diversity [53].

In recent years, tools that may be used for inspection of domain architectural variations between isoforms have been developed. First is the ASPicDB (Alternative Splicing Prediction Database) [35••] which contains annotations of the alternative splicing pattern of human genes as well as functional annotation of the predicted isoforms, including protein domain assignments. Second, Salomonis *et al.* have developed AltAnalyze [62], a tool for RNAseq and microarray analysis, where domain graphs are included in the multiplatform package.

Concluding remarks and future outlook

The main challenge for accurate assessment of the importance of alternative splicing for domain architectural changes is improved identification of functional isoforms at the protein level. As stated above, there are mainly two approaches that have been used to attempt to achieve this: use of evolutionarily conserved patterns or direct studies of the protein isoforms. Assuming that the recent observations of rapidly evolving changes in isoforms between species is correct [63•,64•] many isoforms that are not conserved may still be functional.

Given the limitations of using evolutionary conservation and the abundance of apparently 'nonfunctional' transcripts, it is clear that high throughput proteomics will play an important role in further elucidating the alternative isoforms that are expressed at the protein level [65•].

Clearly, we are only now beginning to understand the function and scope of alternative splicing and it might therefore be too early to definitively give an answer to the question of how it affects domain architecture. However, even from the possibly rather limited data available today it seems like repeated domains and intrinsically disordered regions [42•] are over-represented in alternative spliced isoforms [51•].

Acknowledgements

This work was supported by grants from the Swedish Research Council (VR-NT 2009-5072 and VR-M 2010-3555), SSF, the Foundation for Strategic Research, Science for Life Laboratory; the EU 7th Framework through the EDICT project, contract no: FP7-HEALTH-F4-2007-201924. Funding for SL was provided by BILS, Bioinformatics Infrastructure for Life Science.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Rossmann MG, Moras D, Olsen KW: **Chemical and biological evolution of a nucleotide-binding protein**. *Nature* 1974, **250**:194-199.
2. Sonnhammer E, Eddy S, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments**. *Proteins: Struct Funct Genet* 1997, **28**:405-420.
3. Murzin A, Brenner S, Hubbard T, Chothia C: **Scop: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.
4. Orengo C, Michi A, Jones S, Jones D, Swindels MB, Thornton J: **Cath — a hierarchical classification of protein domain structures**. *Structure* 1997, **5**:1093-1108.
5. Elofsson A, Sonnhammer ELL: **A comparison of sequence and structure protein domain families as a basis for structural genomics**. *Bioinformatics* 1999, **15**:480-500.
6. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes**. *J Mol Biol* 2001, **310**:311-325.
7. Ekman D, Björklund AK, Frey-Skott J, Elofsson A: **Multi-domain proteins in the three kingdoms of like — orphan domains and other unassigned regions**. *J Mol Biol* 2005, **348**:231-243.
8. Björklund AK, Ekman D, Elofsson A: **Expansion of protein domain repeats**. *PLoS Comp Biol* 2006, **2**:e114.
9. Ekman D, Björklund AK, Elofsson A: **Quantification of the elevated rate of domain rearrangements in metazoa**. *J Mol Biol* 2007, **372**:1337-1348.
10. Liu M, Walch H, Wu S, Grigoriev A: **Significant expansion of exon-bordering protein domains during animal proteome evolution**. *Nucleic Acids Res* 2005, **33**:95-105 http://dx.doi.org/10.1093/nar/gki152.
11. Nagy A, Szláma G, Szarka E, Trexler M, Bányai L, Patthy L: **Reassessing domain architecture evolution of metazoan proteins: major impact of gene prediction errors**. *Genes* 2011, **2**:449-501.
12. Nagy A, Bányai L, Patthy L: **Reassessing domain architecture evolution of metazoan proteins: major impact of errors caused by confusing paralogs and epaktologs**. *Genes* 2011, **2**:516-561.
13. Hawkins J: **A survey on intron and exon lengths**. *Nucleic Acids Res* 1988, **16**:9893-9908.
14. Gilbert W, Marchionni M, McKnight G: **On the antiquity of introns**. *Cell* 1986, **46**:151-153.
15. Rogozin I, Carmel L, Csuros M, Koonin E: **Origin and evolution of spliceosomal introns**. *Biol Direct* 2012, **7**:11.
16. Gilbert W: **Why genes in pieces?** *Nature* 1978, **271**:501.
17. Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov J, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin J, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston R, Wilson R, Hillier L, McPherson J, Marra M, Mardis E, Fulton L, Chinwalla A, Pepin K,

Gish W, Chissoe S, Wendl M, Delehaunty K, Miner T, Delehaunty A, Kramer J, Cook L, Fulton R, Johnson D, Minx P, Clifton S, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs R, Muzny D, Scherer S, Bouck J, Sodergren E, Worley K, Rives C, Gorrell J, Metzker M, Naylor S, Kucherlapati R, Nelson D, Weinstock G, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith D, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee H, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis R, Federspiel N, Abola A, Proctor M, Myers R, Schmutz J, Dickson M, Grimwood J, Cox D, Olson M, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans G, Athanasiou M, Schultz R, Roe B, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie W, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey J, Bateman A, Batzoglou S, Birney E, Bork P, Brown D, Burge C, Cerutti L, Chen H, Church D, Clamp M, Copley R, Doerks T, Eddy S, Eichler E, Furey T, Galagan J, Gilbert J, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson L, Jones T, Kasif S, Kaspryzk A, Kennedy S, Kent W, Kitts P, Koonin E, Korf I, Kulp D, Lancet D, Lowe T, McLysaght A, Mikkelsen T, Moran J, Mulder N, Pollara V, Ponting C, Schuler G, Schultz J, Slater G, Smit A, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf Y, Wolfe K, Yang S, Yeh R, Collins F, Guyer M, Peterson J, Felsenfeld A, Wetterstrand K, Patrinos A, Morgan M, de Jong P, Catanese J, Osoegawa K, Shizuya H, Choi S, Chen Y, International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921 http://dx.doi.org/10.1038/35057062.

18. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity**. *Nat Genet* 2002, **30**:29-30.

19. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes**. *Nucleic Acids Res* 2007, **35**:125-131 http://dx.doi.org/10.1093/nar/gkl924.

20. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S: **Function of alternative splicing**. *Gene* 2013, **514**:1-30 http://dx.doi.org/10.1016/j.gene.2012.07.083.

21. Nilsen T, Graveley B: **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 2010, **463**:457-463 http://dx.doi.org/10.1038/nature08909.

22. Harrow J, Denoeud F, Frankish A, Reymond A, Chen C-K, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R: **Gencode: producing a reference annotation for encode**. *Genome Biol* 2006, **7(Suppl. 1)**:S4.1-S4.9 http://dx.doi.org/10.1186/gb-2006-7-s1-s4.

23. Kim E, Goren A, Ast G: **Alternative splicing: current perspectives**. *Bioessays* 2008, **30**:38-47 http://dx.doi.org/10.1002/bies.20692.

24. • Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 2008, **40**:1413-1415 http://dx.doi.org/10.1038/ng.259.
One of the first large scale RNAseq studies focused on alternative splicing in the human genome. They find that approximately 95% of the genes are alternative splices, but do not address the functionality of these.

25. •• Tress M, Martelli P, Frankish A, Reeves G, Wesselink J, Yeats C, Olason P, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski R, Lopez G, Sadowski M, Watson J, Fariselli P, Rossi I, Nagy A, Kai W, Storling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramirez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis S, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones D, Lengauer T, Orengo C, Patthy L, Thornton J, Tramontano A, Valencia A: **The implications of alternative splicing in the ENCODE protein complement**. *Proc Natl Acad Sci U S A* 2007, **104**:5495-5500.
One of the first large-scale studies highlighting that a large fraction of splicing does not appear to produce functional proteins. The examples showing how the structure changes of protein products are very telling for any structural biologist. It is clear that these proteins are not functional.

26. Sultan M, Schulz M, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo M: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. *Science* 2008, **321**:956-960 http://dx.doi.org/10.1126/science.1160342.

27. Hiller D, Jiang H, Xu W, Wong W: **Identifiability of isoform deconvolution from junction arrays and RNA-seq**. *Bioinformatics* 2009, **25**:3056-3059 http://dx.doi.org/10.1093/bioinformatics/btp544.

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

29. Flicek P, Ahmed I, Amode M, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari A, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren W, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat H, Ritchie G, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder S, Aken B, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard T, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle S: **Ensembl 2013**. *Nucleic Acids Res* 2013, **41**:D48-D55.

30. Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database**. *Nucleic Acids Res* 2008, **36(Database issue)**:D753-D760 http://dx.doi.org/10.1093/nar/gkm987.

31. • Frankish A, Mudge J, Thomas M, Harrow J: **The importance of identifying alternative splicing in vertebrate genome annotation**. *Database (Oxford)* 2012, **2012**:bas014.
Database with manual annotation of vertebrate genomes using a cautious approach to make a decision on the functional potential of each splice form. On average they find 6.3 splice forms per human multi exon gene.

32. Pontius JU, Wagner L, Schuler GD: **UniGene: a unified view of the transcriptome**. *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information; 2003.

33. Magrane M, Consortium U: **Uniprot knowledgebase: a hub of integrated protein data**. *Database (Oxford)* 2011, **2011**:bar009.

34. Harrow J, Frankish A, Gonzalez J, Tapanari E, Diekhans M, Kokocinski F, Aken B, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez J, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard T: **GENCODE: the reference human genome annotation for the ENCODE project**. *Genome Res* 2012, **22**:1760-1774 http://dx.doi.org/10.1101/gr.135350.111.

35. •• Martelli P, D'Antonio M, Bonizzoni P, Castrignano T, D'Erchia A, D'Onorio De Meo P, Fariselli P, Finelli M, Licciulli F, Mangiulli M, Mignone F, Pavesi G, Picardi E, Rizzi R, Rossi I, Valletti A, Zauli A, Zambelli F, Casadio R, Pesole G: **ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing**. *Nucleic Acids Res* 2011, **39(Database issue)**:D80-D85.
The only specific alternative splicing database that appears to be up to date. The database is very easy to use and provides an opportunity to download large datasets in various formats. Data on number of transcripts and domain annotations are readily available.

36. • Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, Valencia A, Tress ML: **Appris: annotation of principal and alternative splice isoforms**. *Nucleic Acids Res* 2012, **41**:D110-D117.
A tool to evaluate the probability of splice forms to be functional.

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

38. Melamud E, Moult J: **Stochastic noise in splicing machinery**. *Nucleic Acids Res* 2009, **37**:4873-4886.

39. Tress M, Wesselink J, Frankish A, Lopez G, Goldman N, Loytynoja A, Massingham T, Pardi F, Whelan S, Harrow J, Valencia A: **Determination and validation of principal gene products**. *Bioinformatics* 2008, **24**:11-17.

40. Lopez G, Valencia A, Tress M: **firestar-Prediction of functionally important residues using structural templates and alignment reliability**. *Nucleic Acids Res* 2007, **35(Web Server issue)**: W573-W577.

41. Tress M, Grana O, Valencia A: **SQUARE-determining reliable regions in sequence alignments**. *Bioinformatics* 2004, **20**:974-975.

42. • Hegyi H, Kalmar L, Horvath T, Tompa P: **Verification of alternative splicing variants based on domain integrity truncation length and intrinsic protein disorder**. *Nucleic Acids Res* 2011, **39**:1208-1219.
A recent study of alternative splicing, based on several databases. Detecting a set 505 of high quality spliced variants that appear to be functional. These are enriched in intrinsically disordered protein regions.

43. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L: **Identification and correction of abnormal incomplete and mispredicted proteins in public databases**. *BMC Bioinformatics* 2008, **9**:353 http://dx.doi.org/10.1186/1471-2105-9-353.

44. Leoni G, Le Pera L, Ferre F, Raimondo D, Tramontano A: **Coding potential of the products of alternative splicing in human**. *Genome Biol* 2011, **12**:R9.

45. •• Ezkurdia I, del Pozo A, Frankish A, Rodriguez J, Harrow J, Ashman K, Valencia A, Tress M: **Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function**. *Mol Biol Evol* 2012, **29**:2265-2283.
A paper showing that conserved splice forms between mouse and human are more likely to be functional than non-conserved. The data are obtained from publicly available mass spectrometry studies.

46. Birzele F, Csaba G, Zimmer R: **Alternative splicing and protein structure evolution**. *Nucleic Acids Res* 2008, **36**:550-558.

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

48. Ellis J, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco J, Wang X, Pan Q, O'Hanlon D, Kim P, Wrana J, Blencowe B: **Tissue-specific alternative splicing remodels protein–protein interaction networks**. *Mol Cell* 2012, **46**:884-892 http://dx.doi.org/10.1016/j.molcel.2012.05.037.

49. • Mudge J, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J: **The origins, evolution, and functional potential of alternative splicing in vertebrates**. *Mol Biol Evol* 2011, **28**:2949-2959.
Studying splice forms of 309 conserved protein coding genes between human and mouse Each transcript has been classified to be coding or targeted for nonsense mediated decay.

50. Saltzman A, Kim Y, Pan Q, Fagnani M, Maquat L, Blencowe B: **Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay**. *Mol Cell Biol* 2008, **28**:4320-4330 http://dx.doi.org/10.1128/MCB. 00361-08.

51. • Liu S, Altman RB: **Large scale study of protein domain distribution in the context of alternative splicing**. *Nucleic Acids Res* 2003, **31**:4828-4835.
The paper identifies the domains that are over-represented in alternatively spliced proteins using a curated set of several thousand genes. Domains involved in the processes of cell communication, signaling, development and apoptosis are over-represented.

52. Taneri B, Snyder B, Novoradovsky A, Gaasterland T: **Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific**. *Genome Biol* 2004, **5**:R75.

53. Weatheritt R, Davey N, Gibson T: **Linear motifs confer functional diversity onto splice variants**. *Nucleic Acids Res* 2012, **40**:7123-7131.

54. Ekman D, Light S, Björklund AK, Elofsson A: **What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*?** *Genome Biol* 2006, **7**:R45.

55. Iakoucheva LM, Brown CJ, Lawson JD, Obradovi Z, Dunker AK: **Intrinsic disorder in cell-signaling and cancer-associated proteins**. *J Mol Biol* 2002, **323**:573-584.

56. Buljan M, Chalancon G, Eustermann S, Wagner G, Fuxreiter M, Bateman A, Babu M: **Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks**. *Mol Cell* 2012, **46**:871-883.

57. Nicholson RI, Gee JM, Harper ME: **EGFR and cancer prognosis**. *Eur J Cancer* 2001, **37(Suppl. 4)**:S9-S15.

58. Zanussi S, Doliana R, Segat D, Bonaldo P, Colombatti A: **The human type vi collagen gene. mRNA and protein variants of the alpha 3 chain generated by alternative splicing of an additional 5-end exon**. *J Biol Chem* 1992, **267**:24082-24089.

59. Thorsen K, Sorensen K, Brems-Eskildsen A, Modin C, Gaustadnes M, Hein A, Kruhoffer M, Laurberg S, Borre M, Wang K, Brunak S, Krainer A, Torring N, Dyrskjot L, Andersen C, Orntoft T: **Alternative splicing in colon bladder and prostate cancer identified by exon array analysis**. *Mol Cell Proteomics* 2008, **7**:1214-1224.

60. Zhang P, Greendorfer JS, Jiao J, Kelpke SC, Thompson JA: **Alternatively spliced FGFR-1 isoforms differentially modulate endothelial cell activation of c-YES**. *Arch Biochem Biophys* 2006, **450**:50-62 http://dx.doi.org/10.1016/j.abb.2006.03.017.

61. Garcia J, Gerber SH, Sugita S, Sdhof TC, Rizo J: **A conformational switch in the piccolo C2A domain regulated by alternative splicing**. *Nat Struct Mol Biol* 2004, **11**:45-53 http://dx.doi.org/10.1038/nsmb707.

62. Salomonis N, Nelson B, Vranizan K, Pico A, Hanspers K, Kuchinsky A, Ta L, Mercola M, Conklin B: **Alternative splicing in the differentiation of human embryonic stem cells into cardiac precursors**. *PLoS Comput Biol* 2009, **5**:e1000553.

63. • Merkin J, Russell C, Chen P, Burge C: **Evolutionary dynamics of gene and isoform regulation in mammalian tissues**. *Science* 2012, **338**:1593-1599 http://dx.doi.org/10.1126/science.1228186.
One of the two recent papers highlighting the difference in evolution between gene expression and isoform expression. Isoform expression seems to be more lineage-specific, and conserved alternative exons were identified; widely conserved alternative exons had signatures of binding by MBNL, PTB, RBFOX, STAR, and TIA family splicing factors.

64. • Barbosa-Morais N, Irimia M, Pan Q, Xiong H, Gueroussov S, Lee L, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali C, Wilson M, Kim P, Odom D, Frey B, Blencowe B: **The evolutionary landscape of alternative splicing in vertebrate species**. *Science* 2012, **338**:1587-1593 http://dx.doi.org/10.1126/science.1230612.
One of the two recent papers highlighting the difference in evolution between gene expression and isoform expression. The paper reports significant differences in alternative splicing complexity between vertebrate lineages, with the highest complexity in primates.

65. • Tran J, Zamdborg L, Ahlf D, Lee J, Catherman A, Durbin K, Tipton J, Vellaichamy A, Kellie J, Li M, Wu C, Sweet S, Early B, Siuti N, LeDuc R, Compton P, Thomas P, Kelleher N: **Mapping intact protein isoforms in discovery mode using top-down proteomics**. *Nature* 2011, **480**:254-258 http://dx.doi.org/10.1038/nature10575.
This paper shows that improved proteomics methods can be used to identify splice forms of human proteins using a top-down analysis of whole proteins which has not previously been possible to obtain for such a large set of proteins.

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

67. • Floris M, Raimondo D, Leoni G, Orsini M, Marcatili P, Tramontano A: **MAISTAS: a tool for automatic structural evaluation of alternative splicing products**. *Bioinformatics* 2011, **27**:1625-1629.
A tool to generate structural models of splice forms. Also uses some basic evaluation methods to predict the functionality of the isoforms.