Published in final edited form as:

J Biosci. 2004 September ; 29(3): 261–273.

When proteome meets genome: the alpha helix and the beta strand of proteins are eschewed by mRNA splice junctions and may define the minimal indivisible modules of protein architecture

Sailen Barik

Department of Biochemistry and Molecular Biology (MSB 2370), University of South Alabama, College of Medicine, 307 University Blvd., Mobile, AL 36688-0002, USA (Fax, 251-460-6127; Email, sbarik@jaguar1.usouthal.edu)

Abstract

The significance of the intron-exon structure of genes is a mystery. As eukaryotic proteins are made up of modular functional domains, each exon was suspected to encode some form of module; however, the definition of a module remained vague. Comparison of pre-mRNA splice junctions with the three-dimensional architecture of its protein product from different eukaryotes revealed that the junctions were far less likely to occur inside the α -helices and β -strands of proteins than within the more flexible linker regions ('turns' and 'loops') connecting them. The splice junctions were equally distributed in the different types of linkers and throughout the linker sequence, although a slight preference for the central region of the linker was observed. The avoidance of the α -helix and the β -strand by splice junctions suggests the existence of a selection pressure against their disruption, perhaps underscoring the investment made by nature in building these intricate secondary structures. A corollary is that the helix and the strand are the smallest integral architectural units of a protein and represent the minimal modules in the evolution of protein structure. These results should find use in comparative genomics, designing of cloning strategies, and in the mutual verification of genome sequences with protein structures.

Keywords

Alpha helix; beta strand; exon shuffling; secondary structure; splice junction

1. Introduction

The discovery of 'genes in pieces' nearly three decades ago (Berget et al 1978; Broker et al 1978) prompted a flurry of research about its significance as well as its impact on protein function, evolution, genetic cloning, and recombinant expression. A major conceptual framework was provided by the proposal that each exon encoded a functional unit that could be combined with other exonic units to build more complex proteins from simpler primordial ones (Blake 1978, 1979, 1983; Gilbert 1978, 1987; Go 1981, 1983, 1985; Breitbart and Nadal-Ginard 1986; Duester et al 1986; Go and Nosaka 1987; Patel et al 1987; Shimasaki et al 1988; Hall et al 1989; Tittiger et al 1993; de Souza et al 1996, 1998; Roy et al 1999).

For its obvious evolutionary and biotechnical importance, considerable effort has been devoted in understanding the size and nature of a protein unit encoded in an exon. An initial assumption was that exons encoded independently functioning units, and were shuffled during evolution

(Blake 1978, 1979, 1983; Gilbert 1978, 1987; Go 1981, 1983, 1985). For example, in a study of alcohol dehydrogenase (ADH) and related enzymes (Duester et al 1986) it was noted that the RNA-splice junctions fell at positions that separate the short functional polypeptides believed to comprise the nucleotide-binding domain. As the relationship between structure and function of more proteins became available, however, it was clear that many exons did not code for a module with an independent function. It was also quickly realized – and recent examples are presented in this paper – that many introns, if not most, actually interrupted a functional module.

An important effort in this area was made by Go who, in a series of papers in the eighties (Go 1981, 1983, 1985; Go and Nosaka 1987), proposed that each exon codes for a compact structural unit. The unit was termed a ‘module’ and defined as a compact segment of 28 Å diameter, and about 10–40 amino acids in length. The ‘one exon-one module’ hypothesis held reasonably well in the beginning when applied to a few simple proteins that had a relatively flat 2-dimensional (termed ‘monolayer’ by Go) structure. However, as larger and ‘multilayer’ protein sequences and their structures became available, extremely complex algorithms had to be formulated to define and identify the modules. In fact, within a few years of the original paper, so many exceptions to the Go rule were found that the definition of the module required constant refinement and improvization (de Souza et al 1996, 1998; Go 1983, 1985; Go and Nosaka 1987; Roy et al 1999). Today, the definition of an exonic module remains vague at best. Moreover, most of the early investigators concentrated on a few model proteins such as ADH, globin and triose phosphate isomerase (TPI), partly because their splice junctions were well known at the time. Since then, however, the completion of genome sequencing in many eukaryotes has provided a large number of gene sequences for which such information is available. In this paper, I have carried out a large-scale survey of RNA-splice junctions and their locations on the corresponding proteins. The analysis revealed that the splice junctions tend to avoid the α -helices and the β -strands and that some of the smallest exons code for as little as a single α -helix or β -strand, much smaller than the previously proposed modules.

2. Materials and methods

In each figure the top sequence (#1) had an experimentally determined crystal structure and was, therefore, used as the structural template for other sequences. The NCBI structure accession number for the crystal structures of the human PP1 α isoform (P08129) is 1FJM (Goldberg et al 1995). Those for the human PP1 γ isoform complexed with OA or calyculin A are 1JK7 and 1IT6, respectively. The structure for human PP2B (Kissinger et al 1995) was 1AUI. The splice junctions were obtained from the Ensembl database (www.ensembl.org) search as exemplified by the following example of human PP1 (#1 in figures 1 and 2; NCBI number P08129). A search of the Ensembl site with the key term P08129 led to the Gene ID of ENSG00000172531. The ‘Protein Information’ link on this page led to the PP1 polypeptide sequence that was then chosen to display “Exon Highlighted” (from the default “No Markup” box). In this display, the alternate exons were coloured black and blue on the polypeptide backbone. Based on this, the splice junctions in the figures were marked by red lines. In all figures, the designations of the helices and strands followed standard convention, i.e. α -helices are numbered as αA , αB , αC etc. and β -strands as $\beta 1$, $\beta 2$, $\beta 3$ etc.

In genome-wide studies, the NCBI protein structure database was first searched with keyword of the species (e.g. “human”). All the sequences for which 3D structures were available (e.g. 3212 human sequences as of October 12, 2003) were then selected and the redundant ones manually removed. The rest were retrieved from NCBI RefSeq (a non-redundant database of DNA, mRNA, and protein sequences for all genes) (NCBI handbook, 2002) in XML and FASTA formats. The NCBI files for each gene were processed for the SecStr fields of type ‘helix’ or ‘strand’ using the Apache Xerces Java 2 XML parsing tool, available at

<http://xml.apache.org>. The FASTA files were queried by BLAST against the Ensembl (www.ensembl.org) peptide database of the appropriate species. The highest scoring hit was accessed through the ENSP# link and the peptide sequence was viewed for Exon Boundaries (i.e. RNA splice junctions). The exon start and end positions were selected from the MartView filter of the corresponding protein as a comma-separated text file. The splice junctions were aligned against the structured regions using a custom Java tool. The junctions were initially classified as avoiding or dissecting the structures (helices and strands). Those dissecting the structures were further divided on the basis of whether they were one, two or three amino acid residues into the structure. To check the Ensembl database for accuracy, the intron-exon boundaries of randomly chosen genes (representing roughly 1% of the total sequences presented in table 1) were retrieved manually from the corresponding sequences in NCBI RefSeq by comparing the DNA and mRNA sequences. Random sampling was also performed to check the batch results with the 'Exon Highlighted' feature described above. No discrepancies were observed. In some of the subsequent analyses a different retrieval and archival software was used (Bussow et al 2002) with essentially similar results.

Multiple alignments of the protein sequences (see figures 1, 3) were performed using the ClustalW program at the European Bioinformatics Institute (EMBL) server, and later refined by visual inspection and manual colouration. The statistical significance (P value) was calculated by binomial distribution or by a normal distribution approximating the binomial, as appropriate. It should be pointed out that for a number of phosphatase sequences presented here, the intron-exon structures are predicted. However, this should not have a major impact on the final conclusions of the paper for the following reasons. The accuracy of the prediction algorithms have significantly improved recently such that the predicted mRNA sequence match very well with experimentally determined cDNA or EST sequences. Curated databases such as RefSeq authenticate the predictions by sequence homology within conserved protein families as well as across the phylogenetic tree. Thus, the extreme sequence homology among PP1 virtually guarantees that all the PP1 sequences presented in figure 1 were correct because they all matched the experimentally determined sequence of PP1#1. In any case, the global dataset of table 1 is not affected by prediction errors, because only those proteins for which the crystal structures have been solved were retrieved for further analysis. Therefore, their primary sequences were definitively established, and when compared with the genomic sequence, produced the exact introexon boundaries, i.e. splice junctions.

3. Results

To start with, representative examples of proteins families – specifically, the Ser/Thr protein phosphatases of the PP1, PP2B, and PP2C families – were chosen for analysis. Because the members of a family or superfamily of proteins are highly conserved in sequence and function, the knowledge of the 3D structure of one member allows extension of the structure to the other members. PP1, in particular, is the most conserved phosphatase in biology (Kumar et al 2002). The amino acid sequence identities among the eleven PP1 polypeptides shown in figure 1 ranged from 69% to 88%. In pairwise examples, the sequence identity of PP1#1 with PP1#2, PP1#3 and PP1#4 (figure 1) were 84%, 87% and 84% respectively. When conservative replacements were included the similarities were expectedly higher; the similarity between PP1#1 and #6, for instance, soared up to 93%. Where the three-dimensional structures of two members were available, they were indeed found to be identical, such as the crystal structures of human PP1 α and PP1 γ isoforms (NCBI structure accession numbers 1FJM and 1JK7, respectively). Such similarities allowed us to correlate the exonic configurations with functional architecture in a large number of PP1 orthologs and paralogs.

As members of the PPP superfamily, the PP1 class also shares significant sequence similarity with the PP2B class within the central 'catalytic core' (Barton et al 1994; Ansai et al 1996)

that includes multiple 'signature motifs'. Crystallographic and mutational analyses have identified the critical role of specific amino acid residues within these motifs in metal ion-binding, interaction with the substrate phosphate group and co-ordination with water molecules (Das et al 1996; Egloff et al 1995; Goldberg et al 1995; Kissinger et al 1995; Voegtli 2000). Extra sequences outside the catalytic core are more diverse in nature and participate in nonenzymatic functions such as interaction with accessory proteins, regulatory subunits, and inhibitors (Zhuo et al 1994; Wera and Hemmings 1995; Ansai 1996; Zhang and Lee 1997; Barford et al 1998; Connor et al 1998, 1999; Watanabe et al 2001).

The catalytic domain of these phosphatases can be divided broadly into two tightly linked sub-structures (Das et al 1996; Egloff et al 1995; Goldberg et al 1995; Kissinger et al 1995; Voegtli et al 2000). The N-terminal region (the 'N-subdomain', approximately 182 residues long) includes a β - α - β - α - β metal-coordinating unit (β 2 through β 4 in figures 2, 4) and forms a compact structure in which two metal ions are embedded. The C-terminal region (the 'C-subdomain', approximate residues 183–300) forms an irregular structure that sits on top of the N-subdomain and provides potential binding sites for substrates and inhibitors. Thus, all of the N-subdomain and about 90 residues of the C-subdomain, a total of about 270 residues, comprise the conserved 'catalytic core', whereas the remaining portion of the C-subdomain is involved in regulation, and thus, exhibit greater sequence variability.

PP2C possesses a very different primary structure but similar tertiary structure (Das et al 1996), as graphically indicated in figure 5. Like the PPP enzymes (e.g. PP1 in figures 1, 2 and PP2B in figures 3, 4), it also contains a binuclear metal-binding site bridging two β -strands within a similar β -sandwich architecture. The overall arrangement of helices and strands is also comparable to the PPP phosphatases (figure 5). Moreover, amino acid residues shown to be important in phosphate binding and catalysis are also similarly placed, a few examples of which are marked by coloured dots (figure 5).

For the PP1 and PP2B families, multiple sequence alignments (figures 1, 3) are shown to illustrate the similarity and the amino acid sequences that make up the various domains and structures. For these sequences and for PP2C, a graphic representation (figures 2, 4, 5) is also provided for a bird's eye view of the exonic pattern and how it varies from member to member in the same family. Studies of two other enzyme families, viz., dual-specificity phosphatase and glycosylasparaginase led to essentially similar conclusions, and therefore, their results are not shown.

In the second part, we moved beyond individual protein families and performed a large-scale analysis of 4267 proteins from five eukaryotic species, namely human, mouse, rat, *Drosophila* and *C. elegans* (table 1). These sequences covered essentially all eukaryotic proteins for which three-dimensional structure and gene structure are currently available. The larger data set essentially extended and vindicated the family-specific data. The general rules emanating from all these analyses are summarized below.

3.1 The RNA-splice junctions rarely dissect protein helices and strands

The most basic observation was that the splice junctions (red lines in the figures) were rarely located within the α -helices and the β -strands (the double-headed arrows and boxes in the figures); instead they were mainly in the flexible loop regions that connected the helices and strands. The phosphatase sequences can be used to illustrate how this conclusion was arrived at. In figure 1, for example, PP1 sequence #1 (NCBI protein P08129) contains 330 amino acid residues and 6 RNA splice junctions obtained from Ensembl as described in § 2. Denoted by red lines (figures 1, 2), these junctions are located between α A' and β 1', β 2 and α B, etc. The helices and strands (denoted by the double-headed arrows and rectangles in figure 1) cover a total of 158 residues, which is 48% of the total length of the protein. Therefore, by random

chance, about half of the RNA-splice junctions should fall outside the structured areas of the protein. In reality, however, 5 junctions out of 6 fell outside (or barely inside, in case of $\beta 14$) of all structures ($P < 0.0199$). The lone exception – the junction that fell squarely within the αE helix in PP1 #1, #6 and #7 (figures 1, 2) – is discussed later. As seen in all the figures presented in this paper, essentially similar conclusions could be drawn from the other PP1 members or the PP2B and PP2C family phosphatases.

Results of the expanded study of 4267 proteins are shown in table 1, which also include one phosphatase from each class presented in the figures. In this final tally, 72–82% of all RNA-splice junctions were completely outside of the two structures (helix and strand) despite the fact that only about 44–51% of the protein length (amino acid residues) was unstructured. The statistical significance of this can be demonstrated by using the exact numbers from a single species. For example, in humans, a total of 3212 proteins were analysed, and in all these proteins taken together, 82% of the splice junctions fell within 51% of the protein sequence that was unstructured (table 1), demonstrating the preference of junctions for the unstructured regions. Among the splice junctions that did fall inside a protein secondary structure, the majority barely grazed the structure, being located within 1 or 2 amino acids of one end of the structure. An example is the junction that is located just one residue into the C-terminal side of the $\beta 14$ strand (SFQ/ILK) in PP1 # 2, #4, #6, #7 and #10. When the three kinds of splice junctions – outside and barely inside by 1 or 2 residues – were added in humans, 97% splice junctions obeyed the secondary structure avoidance rule. Splice junctions that were further inside the structures (i.e. 3 or more amino acids into the structure) were extremely rare, ranging only 3–7% in all five species. To restate, although about half of the protein length was structured, only about 1 in 20 splice junctions of RNA fell deep inside the protein structures (table 1). The statistical probability (P) of these numbers to happen by random chance was so close to zero that it could not be mathematically expressed here. Lastly, in the discussion of the three-dimensional structures later, it will be seen that the majority of these apparent exceptions occurred in structures that were far away from the catalytic pocket. Thus, it is fair to conclude that these exceptions further prove the rule.

3.2 The locations of RNA splice junctions are conserved more with respect to protein structure than amino acid or nucleotide sequences

First, a given splice junction often occurred in multiple members of a family but located within the same two structural elements. In many such cases, however, they were not in the exact same place in the primary structure, usually differing by one or two amino acids. Using PP1 as an example again, the splice junction between $\beta 10$ and $\beta 11$ occurs in six out of ten PP1, and all six occurred precisely between Q and V in the invariant sequence HQVV (figure 1). In contrast, the junction between $\beta 2$ and αB occurred in seven out of ten sequences (#1, #5–10) and within the invariant stretch CGDIHGQ but between different residues. In PP1 #1, #6–8, it is between residues G and D, whereas in #5 it is between H and G, etc. (figure 1). Regardless, neither kind touched $\beta 2$ or αB , suggesting that the exact location of the RNA splice junction in the protein primary structure could be relatively variable so long as it avoided the two fundamental structural elements of the protein.

Second, as shown in figure 5, PP2C also follows the basic rule in that most of the RNA splice junctions (red bars) are outside the protein structures. However, PP2C is unique in that its nucleotide and amino acid sequence are highly different from those of PP1 or PP2B; however, it is very similar to the others in its three-dimensional structure. Clearly, PP2C have evolved independently and not by shuffling sequences with other phosphatases. The validation of the rule in PP2C again suggests a preference of structural landmarks over primary sequence in placing the RNA splice junctions.

Lastly, as a general rule, the nucleotide sequences of the genes within a protein family are generally much more divergent than the protein sequences, in part owing to the degeneracy of the codons. An extreme example is PP1. As shown, the primary structure of PP1 proteins is highly conserved in all species including the malaria parasite, *P. falciparum* (figure 1). For example, the human and the *P. falciparum* PP1 sequences (#1 and #2 in figure 1) are 84% identical at the amino acid level. However, unlike the human genome, the *P. falciparum* genome is extremely AT-rich and this is also reflected in the PP1 coding sequence; thus, the homology between the cDNA sequences of human and *P. falciparum* PP1 (at the nucleotide level) is practically undetectable (not shown). Moreover, although *P. falciparum* splice sites follow the GU-AG rule, there is no detectable similarity between the human and *P. falciparum* PP1 RNA sequences flanking the splice junctions. The same observation applies to other orthologs of human and *P. falciparum* that we have analysed so far (data not shown). It thus appears that the most obvious landmarks in terms of the occurrence of RNA splice junctions are the structural regions of the protein rather than the primary structure of the protein or the nucleotide sequence of the gene or RNA.

3.3 The smallest exons do not encode any recognizable activity

Even a cursory look within any protein family revealed that the exonic arrangement of the domains varied considerably, even when the goal was to produce essentially the same enzyme. For example, in PP1 sequences #9 and #10 (figure 2) one large exon coded for nearly three-quarter of the catalytic domain. In contrast, the catalytic domains of other PP1 enzymes are spread over multiple exons. The smallest exon in this family is found in PP1#6, which codes for a 24-residue peptide encompassing just the helix G plus a few unstructured amino acids on either side (figures 1, 2). In many PP1 sequences another small exon codes for the $\alpha A'$ helix only (figure 2). As discussed below, the functional catalytic pocket requires the participation of a large number of helices and strands drawn from the whole length of the minimal catalytic core. Therefore, these small peptides are obviously incapable of any enzymatic activity.

Deletion analysis of the 221 amino acid-long bacteriophage lambda phosphatase showed that loss of 2–3 amino acids from either terminus resulted in an inactive protein, suggesting that it may represent a minimal phosphatase sequence (Ansai et al 1996; Barik 1993). Taking PP1 as a representative example, the similar region roughly corresponds to residue 48–280 (Ansai et al 1996), and a quick look at figure 1 reveals that there is no single exon that covers this whole stretch. For a more in-depth and mechanistic analysis, the amino acid residues that interact with the cation and/or the substrate were marked by small circles in figures 1 and 2. PP1 has been shown to contain two metal ions in its catalytic pocket (Goldberg et al 1995; Egloff et al 1995). One of them is coordinated by D64, H66, D92 (green circles), as well as oxygens of water and the substrate's phosphate group (the numbers correspond to PP1#1). The other metal is held by D92, N124, H173, H248 (blue circles), as well as by phosphate and water. A number of residues directly coordinate with the phosphate group; these include R96, N124, H125, R221, and Y272 (purple circles). Figure 2 clearly shows that only PP1 #9 and #10 have large exons that are barely big enough to encode all these residues. Even then, their N-termini are too close to D64, and moreover, deletion studies mentioned above showed that this stretch is too short to have phosphatase activity. In all the other PP1, the exons are much too small and thus, encode only a subset of residues that are clearly incapable of forming even a single metal ion pocket. One can make essentially similar observations for PP2B and PP2C (figures 3-5). Clearly, there is no exon in the phosphatase families that is large enough for a discernible functional activity.

3.4 Most exons contain at least one structural unit

Although the small exons did not contain any functional domain, even the smallest ones contained at least one integral α -helix or a β -strand. For instance, the complete helix G in PP1

#6 (figure 2) was contained in one of the smallest exons. This is in fact related to the first observation that RNA splice junctions do not split the structural units, which leads to the corollary that if there is a structural unit in an exon it would be complete and uninterrupted.

3.5 Splice junctions show little preference for the nature or sequence of the protein linkers

It is clear from the foregoing that the majority of splice junctions reside within the relatively less structured peptide sequences that join the helices and strands. For the sake of generalization, these regions are referred to as “linker” in this paper. Such linker sequences are in fact made of different variations of “loops” and “turns” which, in contrast to the helices and strands, are relatively degenerate, flexible and poorly defined (Ponting and Russell 2002). Linkers have a wide spectrum of amino acid sequences, although certain positions may exhibit preferences for Gly, Asp or Asn to conform to the torsional angles permissible in the Ramachandran plot (Pal and Chakrabarti 2002; Sheik et al 2002). Notwithstanding their variability, attempts have been made to classify the linkers based on their conformation and the nature of the secondary structure that they connect (Sheik et al 2002). In simple terms, three fundamental classes of linkers have been proposed: α - α , α - β and β - β . It was, therefore, tested whether the splice junctions display any preference for a particular class.

The 4,267 proteins listed in table 1 contained a total of 19,201 splice junctions, averaging at roughly 4.5 introns per gene, which is consistent with recent estimates from eukaryotic genomes (Garcia-Blanco 2003). Among the splice junctions that were within the linkers, 12,000 were randomly selected and analysed further. As shown in table 2, the three classes of linkers contained comparable numbers of splice junctions without any strong bias for a particular class. Because of the large variety of potential secondary structures in the linkers and the inherent uncertainty of assigning the structures, no attempt was made here to classify the linkers by structure. Nevertheless, a cursory examination of a few linkers of well-known structure revealed no preference of splice junctions for any class (data not shown). It was then tested whether, within a linker, certain regions are preferred. For this purpose, the length of each linker (i.e. number of amino acid residues) was taken as 100, and was then divided into five equal segments (0–20%, 21–40%, 41–60%, 61–80% and 81–100%). A linker with only two residues, for example can only have splice junctions between these two residues, and thus, all splice junctions will belong to the central segment (41–60%) of the linker. At the end, the number of splice junctions found in each segment was counted among all the linkers, and the results are shown in table 2. It is clear that splice junctions do not exhibit any pronounced bias favouring a particular segment of the linker although they seem to be slightly more prevalent towards the center of the linker as compared to the termini.

4. Discussion

In this paper I propose that the helix and the strand are the minimal secondary structures of protein architecture and as such are treated as indivisible modules during structural evolution. Once formed, they evolve as integral units to generate larger domains and folds of recognizable function. For this to be true, the small structural modules would have evolutionary advantages in each step. In other words, the exonic polypeptides must have served a role – even if a primitive one – early in evolution, which may or may not be related to their present-day activity. In fact, a detailed re-examination of four ancient proteins: namely; ADH, globin, pyruvate kinase (PK), TPI by Doolittle (Stoltzfus et al 1994) again clearly showed a lack of exonic correspondence to larger protein modules, such as the ones proposed by Go. Even in these proteins, however, a relative crowding of splice junctions outside of the helices and strands could be discerned. In class I ADH, for example, 13 out of the 20 known junctions were either outside or near the end of helices or strands. Such numbers for PK and TPI were, respectively, 14 out of 16, and 12 out of 14 (Stoltzfus et al 1994). The only apparent exceptions were members

of the globin family (Stoltzfus et al 1994); however, eukaryotic globin functions as a tetramer and, therefore, the individual secondary structural elements may play a less direct role in its function. Thus, it is quite possible that even the ancient exons, encoding only one or a few helices and strands, still provided partial functions that were evolutionarily important. A partial metal-binding domain, for example, could have functioned as an electron donor to a non-metal entity, such as another amino acid residue in the same protein or in a different protein. Alternately, they could have played a structural role, away from the catalytic pocket, to bring the residues of the catalytic pocket together.

As implied previously, the concept of the fundamental units of tertiary structure is being increasingly appreciated. An example is the recently cataloged Structural Classification of Proteins (SCOP) and related databases (Coulson and Moult 2002; Getz et al 2002; Gough and Chothia 2002; Lo Conte et al 2002; Mallika et al 2002; Offer et al 2002) that recognize all protein folds to be made up of various arrangements of α -helices and β -strands, connected by relatively unstructured and flexible loop regions. Each SCOP fold further leads to superfamily, family, protein, and the species in which the proteins are found. The importance of the SCOP classification is in the understanding that apparently unrelated superfamily domains often derive from similar SCOP folds. For example all three glycosyltransferases, metal-dependent hydrolases, and aldolases, share the α/β barrel fold. Analysis presented here suggests that the fundamental unit of evolution of the protein architecture is smaller than a SCOP fold, and may be as small as a single helix or strand.

Why do RNA splice junctions seem to avoid these two fundamental structural units of proteins? Although the exact answer is not known, one can presume that there is a higher evolutionary price for breaking a unit of secondary structure than a region of little or no structure. As an example, let's consider the α -helix which takes up about a quarter of all amino acids in an average protein and plays a crucial role in protein structure and function. The α -helix is a relatively rigid structure with an average pitch of 3.6 amino acids (5.4 Å), and has stringent architectural requirements (Offer et al 2002). Within the helix, every peptide bond participates in H-bonding, with three to four H-bonds connecting two adjacent turns, and thus contributing to the extraordinary stability of the helix. The stringent pitch and twist of the helix ensure that an amino acid side chain makes critical interactions with another side chain three to four residues away on either side of it. Thus, positively charged amino acids are often found three residues away from negatively charged ones, permitting the formation of an ion pair. Similarly, two aromatic residues are often placed three residues apart, allowing a hydrophobic interaction. Moreover, the biological role of the protein often depends on the presence of a set of amino acids on the same face of the protein, such as the face that interacts with DNA, or forms protein-protein dimers. Insertion or deletion of a small number of amino acids within the helix thus distorts the three-dimensional alignment between the residues upstream and downstream of the insertion, resulting in loss of function (Pu and Struhl 1991; Perera et al 2001; Marin et al 2002). As the insertion and excision of introns has the potential for occasional imprecision, it risks disruptions of the helical structure and function. Essentially similar arguments can be invoked for the β -strand that form a zigzag structure and is often arranged in parallel or antiparallel pleated strands such that the loss or gain of a single residue in one β -strand will throw the whole pleat out-of-phase. Thus, intra-helical or intra-strand insertion and deletion of amino acids might have been eliminated during evolution because they often led to inactive proteins that were evolutionary dead ends.

An alternate – but not mutually exclusive – possibility was that the helix and the strand regions simply lacked a specific sequence at the RNA level that might be important for intronic insertions (splice junctions). This seems unlikely on the following grounds. First, the RNA splice junctions have an extremely loose consensus sequence, which is too variable to have a predictive value. In mammals, for example, the splice 'donor' and 'acceptor' sequence of the

major class of introns (U2-dependent) are roughly C(A)AG/GUA(G)AGU and U(C)NNC(U)AG/G, respectively, where N is any nucleotide (Garcia-Blanco 2003), the alternative nucleotides are in parenthesis, and the only truly invariant residues are the terminal GU and AG dinucleotides (underlined) of the intron (Breathnach and Chambon 1981). In fact, such sequences are routinely found within exons, suggesting that splice junctions require additional primary or secondary structural elements of the pre-mRNA that are currently unknown. Second, one might argue that the neighbouring sequences are important in placing the splice junctions and that such sequences are more prevalent in the linkers. If this is true, these sequences should be reflected in the amino acid sequences as well. However, a survey of amino acid residues in the vicinity of the splice junctions did not reveal any consistent residue(s) that were absent in helices and strands (data not shown). For example, Pro and Gly are generally rare in helices but occur in β -turns; likewise, the Val-Pro dyad shows a strong preference for loops (Pal and Chakrabarti 2002; Ponting and Russell 2002). However, even a cursory look at figures 1 and 3 reveal that the splice junctions did not necessarily locate next to Pro or Gly residues or a Val-Pro dyad (data not shown). Thus, the preferential occurrence of splice junctions in the linker regions is not due to a differential primary structure of the RNA or protein.

As indicated in the various figures and in table 1, a few exceptions to this rule were seen whereby splice junctions were found inside a helix or a strand. Again, no amino acid or nucleotide sequence pattern was found to be common among these sites (data not shown). Attention was, therefore, concentrated on those proteins in which the functional role of such structures was known from detailed structure-function analyses. These studies revealed that in almost all such exceptions, the affected helix or strand did not play a direct role in the protein's function. Helix α E in PP1 and PP2B is an example (figures 1, 3). Two other examples are the β 13 strand and the α H helix in PP2B (figure 3). All these structures are far enough from the catalytic pocket such that small changes in their sequence should not appreciably affect the function of the protein. The last two structures mentioned in PP2B are in fact not in the catalytic domain; they are in the C-terminal accessory domain that binds the CnB subunit, and therefore, play a nonessential role in catalysis (Kissinger et al 1995). Thus, intronic insertions within these structures would be relatively tolerable and would persist during evolution. As mentioned earlier, these exceptions further prove the generality of the rule that RNA splice junctions tend to avoid the α -helices and the β -strands of proteins. A slight preference of the splice junctions to be clustered near the center of the linkers (table 2) may simply reflect this trend to stay far from the flanking helix or strand. The lack of preference for any class of linker (table 2) also supports the view that the linker regions in general are structurally less rigid and demanding than the helices or strands that flank them.

In conclusion, comparison of intron-exon arrangement of a gene and the secondary structure of the corresponding protein led to the empirical rule that RNA splice junctions tend to avoid the α -helix and β -strand of proteins. This is most likely because helices and strands are relatively higher order structures and less tolerant of sequence changes. RNA splicing within these sequences during exon shuffling runs the risk of irreversibly damaging these structures, and was thus evolutionarily disfavoured and eliminated.

Acknowledgements

This study was supported in part by NIH grants AI045803 and EY013826, respectively from the National Institute of Allergy and Infectious Diseases (NIAID) and the National Eye Institute (NEI). Sincere thanks are offered to the NCBI and Ensembl staff for allowing free online access to the databanks and the analysis tools. Ensembl is a project of the Sanger Institute, funded by Wellcome Trust. Thanks are also due to Titus Barik for expert assistance in writing the computer programs and to Dr Dulal Bhaumik (University of Illinois at Chicago, USA) for calculation of statistical significance. A preliminary report of these studies was presented in February 2002 in the Physiology Seminar Series, Department of Physiology.

Abbreviations used

ADH, Alcohol dehydrogenase; PK, pyruvate kinase; SCOP, Structural Classification of Proteins; TPI, triose phosphate isomerase.

References

- Ansai T, Dupuy LC, Barik S. Interactions between a minimal protein serine/threonine phosphatase and its phosphopeptide substrate sequence. *J. Biol. Chem* 1996;271:24401–24407. [PubMed: 8798696]
- Barford D, Das AK, Egloff MP. The structure and mechanism of protein phosphatases: insights into catalysis and regulation. *Annu. Rev. Biophys. Biomol. Struct* 1998;27:133–164. [PubMed: 9646865]
- Barik S. Expression and biochemical properties of a protein serine/threonine phosphatase encoded by bacteriophage lambda. *Proc. Natl. Acad. Sci. USA* 1993;90:10633–10637. [PubMed: 8248155]
- Barton GJ, Cohen PT, Barford D. Conservation analysis and structure prediction of the protein serine/threonine phosphatases. Sequence similarity with diadenosine tetraphosphatase from *Escherichia coli* suggests homology to the protein phosphatases. *Eur. J. Biochem* 1994;220:225–227. [PubMed: 8119291]
- Berget SM, Berk AJ, Harrison T, Sharp PA. Spliced segments at the 5' termini of adenovirus-2 late mRNA: a role for heterogeneous nuclear RNA in mammalian cells. *Cold Spring Harb. Symp. Quant. Biol* 1978;42:523–529. [PubMed: 277361]
- Blake CCF. Do genes-in-pieces imply proteins-in-pieces? *Nature (London)* 1978;273:267.
- Blake CCF. Exons encode protein functional units. *Nature (London)* 1979;277:598. [PubMed: 423957]
- Blake CCF. Exons – present from the beginning? *Nature (London)* 1983;306:535–537. [PubMed: 6646232]
- Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. *Annu. Rev. Biochem* 1981;50:349–383. [PubMed: 6791577]
- Breitbart RE, Nadal-Ginard B. Complete nucleotide sequence of the fast skeletal troponin T gene. Alternatively spliced exons exhibit unusual interspecies divergence. *J. Mol. Biol* 1986;188:313–324. [PubMed: 3735424]
- Broker TR, Chow LT, Dunn AR, Gelinis RE, Hassell JA, Klessig DF, Lewis JB, Roberts RJ, Zain BS. Adenovirus-2 messengers – an example of baroque molecular architecture. *Cold Spring Harb. Symp. Quant. Biol* 1978;42:531–553. [PubMed: 277362]
- Bussow K, Hoffmann S, Sievert V. ORFer – retrieval of protein sequences and open reading frames from GenBank and storage into relational databases or text files. *BMC Bioinformatics* 2002;3:40. [PubMed: 12493080]
- Connor JH, Kleeman T, Barik S, Honkanen RE, Shenolikar S. Importance of the $\beta 12$ – $\beta 13$ loop in protein phosphatase-1 catalytic subunit for inhibition by toxins and mammalian protein inhibitors. *J. Biol. Chem* 1999;274:22366–22372. [PubMed: 10428807]
- Connor JH, Quan HQ, Ramaswamy NT, Zhang L, Barik S, Zheng J, Cannon JP, Lee EYC, Shenolikar S. Inhibitor-1 interaction domain that mediates the inhibition of protein phosphatase-1. *J. Biol. Chem* 1998;273:27716–27724. [PubMed: 9765309]
- Coulson AF, Moult J. A unfold, mesofold, and super-fold model of protein fold use. *Proteins* 2002;46:61–71. [PubMed: 11746703]
- Das AK, Helps NR, Cohen PT, Barford D. Crystal structure of the protein serine/threonine phosphatase 2C at 2.0 Å resolution. *EMBO J* 1996;15:6798–6809. [PubMed: 9003755]
- de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W. Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl. Acad. Sci. USA* 1996;93:14632–14636. [PubMed: 8962105]
- de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA* 1998;95:5094–5099. [PubMed: 9560234]
- Duester G, Jornvall H, Hatfield GW. Intron-dependent evolution of the nucleotide-binding domains within alcohol dehydrogenase and related enzymes. *Nucleic Acids Res* 1986;14:1931–1941. [PubMed: 2938077]

- Egloff MP, Cohen PT, Reinemer P, Barford D. Crystal structure of the catalytic subunit of human protein phosphatase 1 and its complex with tungstate. *J. Mol. Biol* 1995;254:942–959. [PubMed: 7500362]
- Garcia-Blanco MA. Messenger RNA reprogramming by spliceosome-mediated RNA trans-splicing. *J. Clin. Invest* 2003;112:474–480. [PubMed: 12925685]
- Getz G, Vendruscolo M, Sachs D, Domany E. Automated assignment of SCOP and CATH protein structure classifications from FSSP scores. *Proteins* 2002;46:405–415. [PubMed: 11835515]
- Gilbert W. Why genes in pieces? *Nature (London)* 1978;271:501. [PubMed: 622185]
- Gilbert W. The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol* 1987;52:901–905. [PubMed: 2456887]
- Goldberg J, Huang HB, Kwon YG, Greengard P, Nairn AC, Kuriyan J. Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature (London)* 1995;376:745–753. [PubMed: 7651533]
- Go M. Correlation between DNA exonic regions with protein structural units in hemoglobin. *Nature (London)* 1981;291:90–92. [PubMed: 7231530]
- Go M. Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl. Acad. Sci. USA* 1983;80:1964–1968. [PubMed: 6572956]
- Go M. Protein structures and split genes. *Adv. Biophys* 1985;19:91–131. [PubMed: 3915183]
- Go M, Nosaka M. Protein architecture and the origin of introns. *Cold Spring Harb. Symp Quant. Biol* 1987;52:915–924. [PubMed: 3454299]
- Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* 2002;30:268–272. [PubMed: 11752312]
- Hall DH, Liu Y, Shub DA. Exon shuffling by recombination between self-splicing introns of bacteriophage T4. *Nature (London)* 1989;340:575–576. [PubMed: 2770862]
- Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW, et al. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. *Nature (London)* 1995;378:641–644. [PubMed: 8524402]
- Kumar R, Adams B, Oldenburg A, Musiyenko A, Barik S. Characterisation and expression of a PP1 serine/threonine protein phosphatase (PfPP1) from the malaria parasite, *Plasmodium falciparum*: demonstration of its essential role using RNA interference. *Malar. J* 2002;1:5. [PubMed: 12057017]
- Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 2002;30:264–267. [PubMed: 11752311]
- Mallika V, Bhaduri A, Sowdhamini R. PASS2: a semiautomated database of protein alignments organized as structural superfamilies. *Nucleic Acids Res* 2002;30:284–288. [PubMed: 11752316]
- Marin EP, Krishna AG, Sakmar TP. Disruption of the alpha5 helix of transducin Impairs rhodopsin-catalyzed nucleotide exchange. *Biochemistry* 2002;41:6988–6994. [PubMed: 12033931]
- Offer G, Hicks MR, Woolfson DN. Generalized Crick equations for modeling noncanonical coiled coils. *J. Struct. Biol* 2002;137:41–53. [PubMed: 12064932]
- Pal D, Chakrabarti P. On residues in the disallowed region of the Ramachandran map. *Biopolymers* 2002;63:195–206. [PubMed: 11787007]
- Patel RS, Odermatt E, Schwarzbauer JE, Hynes RO. Organization of the fibronectin gene provides evidence for exon shuffling during evolution. *EMBO J* 1987;6:2565–2572. [PubMed: 3119323]
- Perera R, Owen KE, Tellinghuisen TL, Goralenya AE, Kuhn RJ. Alphavirus nucleocapsid protein contains a putative coiled coil alpha-helix important for core assembly. *J. Virol* 2001;75:1–10. [PubMed: 11119567]
- Ponting CP, Russell RR. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct* 2002;31:45–71. [PubMed: 11988462]
- Pu WT, Struhl K. The leucine zipper symmetrically positions the adjacent basic regions for specific DNA binding. *Proc. Natl. Acad. Sci. USA* 1991;88:6901–6905. [PubMed: 1871104]
- Roy SW, Nosaka M, de Souza SJ, Gilbert W. Centripetal modules and ancient introns. *Gene* 1999;238:85–91. [PubMed: 10570987]
- Sheik SS, Sundararajan P, Hussain AS, Sekar K. Ramachandran plot on the web. *Bioinformatics* 2002;18:1548–1549. [PubMed: 12424132]

- Shimasaki S, Koga M, Esch F, Cooksey K, Mercado M, Koba A, Ueno N, Ying SY, Ling N, Guillemin R. Primary structure of the human follistatin precursor and its genomic organization. *Proc. Natl. Acad. Sci. USA* 1988;85:4218–4222. [PubMed: 3380788]
- Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF. Testing the exon theory of genes: the evidence from protein structure. *Science* 1994;265:202–207. [PubMed: 8023140]
- The NCBI handbook [Internet]. National Library of Medicine (US), National Center for Biotechnology Information; Bethesda (MD): Oct.. 2002 RefSeq is on Chapter 17. Available from <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books>
- Tittiger C, Whyard S, Walker VK. A novel intron site in the triosephosphate isomerase gene from the mosquito *Culex tarsalis*. *Nature (London)* 1993;361:470–472. [PubMed: 8429888]
- Voegtli WC, White DJ, Reiter NJ, Rusnak F, Rosenzweig AC. Structure of the bacteriophage lambda Ser/Thr protein phosphatase with sulfate ion bound in two coordination modes. *Biochemistry* 2000;39:15365–15374. [PubMed: 11112522]
- Watanabe T, Huang HB, Horiuchi A, da Cruze Silva EF, Hsieh-Wilson L, Allen PB, Shenolikar S, Greengard P, Nairn AC. Protein phosphatase 1 regulation by inhibitors and targeting subunits. *Proc. Natl. Acad. Sci. USA* 2001;98:3080–3085. [PubMed: 11248035]
- Wera S, Hemmings BA. Serine/threonine protein phosphatases. *Biochem. J* 1995;311:17–29. [PubMed: 7575450]
- Zhang L, Lee EY. Mutational analysis of substrate recognition by protein phosphatase 1. *Biochemistry* 1997;36:8209–8214. [PubMed: 9204865]
- Zhuo S, Clemens JC, Stone RL, Dixon JE. Mutational analysis of a Ser/Thr phosphatase. Identification of residues important in phosphoesterase substrate binding and catalysis. *J. Biol. Chem* 1994;269:26234–26238. [PubMed: 7929339]

Figure 1.

RNA-splice junctions versus protein structures in the PP1 family. The PP1 sequences are designated #1 through #10; their accession numbers are: #1 (P08129; *H. sapiens* PP1α); #2 (AAM54063; *Plasmodium falciparum*); #3 (P48461; *D. melanogaster* PP1α); #4 (P48462; *D. melanogaster* PP1β); #5 (O82734; *A. thaliana* PP1 isozyme 8); #6 (AH003265, AAA85093; *H. sapiens* PP1β); #7 (AAK18957; *C. elegans*); #8 (BAA97417; *A. thaliana*); #9 (M71715, AAB59322; *S. cerevisiae* GLC7); #10 (M27069; *S. cerevisiae* sds21). The splice junctions (red vertical lines) and α-helices (double arrowheads) and β-strands (boxes) were obtained as described in §2. Amino acid residues involved in binding specific entities are marked by coloured dots: metal ion M1 (green); metal ion M2 (blue); and the substrate phosphate group

(purple). When there are extra amino acids at the C-terminal end that are not shown, this is indicated by dots.

Figure 2.
Graphic view of RNA splice junctions and protein secondary structures in the PP1 family.
Sequences and notations are the same as in figure 1. The catalytic domain is so demarcated.

Figure 3.

RNA splice junctions versus protein structures in the PP2B (calcineurin, CnA) family. The legend is essentially as in figure 1. The sequences are: #1 (NP_066955.1; human PP2B β), #2 (P48454, human PP2B β), #3 (BAA06081.1; *S. pombe*), #4 (Q08209; human PP2B β). The helices and strands are based on the crystal structure of #1 (Goldberg et al 1995). As in PP1 (figure 1), the same residues are involved in binding two metal ions (blue and green dots) and the substrate phosphate group (purple dot). When two amino acids are in lowercase, it means that the sequence between them is not shown for the sake of alignment. In such cases, the removed sequence did not contain any secondary structure.

Figure 4.

Graphic view of RNA splice junctions and protein secondary structures in the PP2B (CnA) family. Sequences are the same as in figure 3, and the legend is similar to figure 2. In addition, domains that bind the calcineurin B subunit (CnB) and the Ca^{+2} -binding subunit, calmodulin (CaM), are so marked.

Figure 5.

Graphic view of RNA splice junctions and protein secondary structures in the PP2C family. The following PP2C sequences are presented and numbered as follows: 1 (NP_187748.1), 2 (NP_187278.1), 3 (AAA67321.1), 4 (CAA86456.2), 5 (AAF47747.1), 6 (AAF56905.1), 7 (NP_187868.1), 8 (NP_002698.1), 9 (NP_002697.1). The legend is similar to figure 2 and the same colour codes have been used to mark some of the residues involved in binding the two metal ions (blue and green dots) and the substrate phosphate group (purple dot). PP2C sequences are inherently more variable than other phosphatases, and this is also reflected in the protein architecture, especially at the two termini.

Genome-wide data showing prevalence of splice junctions outside protein secondary structures^(a).

Table 1

Species ^(b)	No. of proteins analysed ^(c)	Amino acids outside Sec. Str. ^(d) (%)	Splice junction locations relative to Sec. Str. (%)		
			Outside ^(e)	Inside ^(f)	No. of amino acids of Str. 2 3
Human	3212	51	82	10	5
Mouse	526	49	75	10	8
Rat	425	48	74	11	10
<i>Drosophila</i>	84	51	72	12	9
<i>C. elegans</i>	20	44	80	10	6

^(a)“Secondary structures” (Sec. Str.) is defined as the sum of α -helices and β -strands, and were retrieved as described in § 2.

^(b) Only the major eukaryote species were taken for which a substantial number of protein 3D structures were available. For example, although a large number of annotated protein sequences are known for zebrafish, 3D structure has been solved for just one.

^(c) Overlapping peptides and mutant proteins were eliminated from the study.

^(d) Number of amino acid residues present outside the α -helices and β -strands (i.e. present in unstructured linker regions) were counted and expressed as percentage of the total number of amino acids.

^(e) RNA splice junctions that fell outside the secondary structures (i.e. within linkers) were counted and expressed as percentage of total junctions.

^(f) RNA splice junctions that were located inside secondary structures were counted and divided into three groups based on how deep inside the structure they were located: (i) those that were located only one amino acid inside the structure (i.e. barely grazing the structure); (ii) those that were located only two amino acids inside the structure; and (iii) those that were located three or more amino acids inside the structure. The last group is in the greatest violation of the proposed rule, but represents only a small percentage of the total introns. All % numbers were rounded to the nearest integer.

Table 2

Location of splice junction within various types of protein linkers^(a).

Linker type	Location of splice junctions within linker length (percent length of linker, from the N-terminal end)					Row total
	1–20%	21–40%	41–60%	61–80%	81–100%	
α -linker- α	596	788	913	853	745	3895
α -linker- β	687	776	995	786	738	3982
β -linker- β	800	879	919	802	723	4123
Column total:	2083	2443	2827	2441	2206	12000
Percent of grand total:	17%	20%	24%	21%	18%	

^(a) A linker was defined as any stretch of protein regardless of conformation that connected two helices, a helix and a strand, or two strands (see § 3.5). The 12,000 linkers were divided into these three types on the left column. The row totals reveal that they are nearly equal in number (3895, 3982, 4123). The linker can be imagined as a straight line, and its full length (100%) is arbitrarily divided into five equal segments (1–20%, 21–40% etc.). Note that in all types of linkers the splice junctions tend to be slightly more clustered in the central segment of the linkers (24%) as opposed to the N-terminus (17%) and C-terminus (18%) of the linkers. Details are given in § 3.5.