The Difference in Structural States between Canonical Proteins and Their Isoforms Established by Proteome-Wide Bioinformatics Analysis

Zarifa Osmanli, Theo Falgarone, Turkan Samadova, Gudrun Aldrian, Jeremy Leclercq, Ilham Shahmuradov, Andrey V Kajava

biomolecules

MDPI

The difference of structural states between canonical proteins and their isoforms established by proteome-wide bioinformatics analysis

Zarifa Osmanli [1,2], Theo Falgarone[1], Turkan Samadova[2], Gudrun Aldrian[1], Jeremy Leclercq[1], Ilham Shahmuradov[2] and Andrey V Kajava[1*]

[1] CRBM, Université de Montpellier, CNRS, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France; andrey.kajava@crbm.cnrs.fr
[2] Institute of Biophysics, ANAS, Baku, Azerbaijan

* Correspondence: andrey.kajava@crbm.cnrs.fr

Abstract: Alternative splicing is an important mean of generating the protein diversity necessary for cellular functions. Hence, there is a growing interest in assessing the structural and functional impact of alternative protein isoforms. Typically, experimental studies are used to determine the structures of the canonical proteins ignoring the other isoforms. Therefore, there is still a large gap between abundant sequence information and meager structural data on these isoforms. During the last decade, significant progress has been achieved in the development of bioinformatics tools for structural and functional annotations of proteins. Moreover, the appearance of the AlphaFold program opened up the possibility to model a large number of high-confidence structures of the isoforms. In this study, using state-of-the-art tools, we performed *in silico* analysis of 58 eukaryotic proteomes. The evaluated structural states included structured domains, intrinsically disordered regions, aggregation-prone regions and tandem repeats. Among other things, we found that the isoforms have less signal peptides, transmembrane regions or tandem repeat regions in comparison with their canonical counterparts. This could change protein function and/or cellular localization. The AlphaFold modelling demonstrated that frequently isoforms, having differences with the canonical sequences, still can fold in similar structures though with significant structural rearrangements which can lead to changes of their functions. Based on the modelling, we suggested classification of the structural differences between canonical proteins and isoforms. Altogether, we can conclude that a majority of isoforms, similarly to the canonical proteins are under selective pressure for the functional roles.

Keywords: isoform, large-scale analysis, protein structure, AlphaFold, canonical protein

1. Introduction

Alternative splicing is one of the principal sources for structural and functional diversity in the proteomes of multicellular organisms. It is a process, which may include or exclude particular exons of a multi-exonic gene from its processed messenger RNA. Different combinations of exons can produce multiple mRNA isoforms of a single gene. It is estimated that up to 95% of human multi-exonic genes are alternatively spliced [1-2]. The average number of the splice variants per human gene is equal to four [3]. All this can drastically increase the number of different proteins in the proteome. Today, most genome-wide information about alternative splicing is generated on the nucleic acid level thanks to high-throughput data such as expressed sequence tags (ESTs) [4], microarrays [5] and RNA-seq data [6]. However, not all splicing variants are expressed as functional proteins. Although a very large number of alternatively spliced variants are detected in RNA-seq studies, large-scale mass spectrometry-based proteomics analyses detect only a small fraction of alternative isoforms on the protein level [7]. One of today's problems in

this area is to establish the real number of splice variants that appear as functional proteins for each gene. In addition to the application of genome-wide mass spectrometry analyses, researchers pay special attention to the protein isoforms with the most cross-species conservation and those that are able to maintain protein structure integrity [1, 8-10].

Although the way to obtain the exact set of the real protein variants may take some time, the data already available thanks to combination of approaches (proteomics, cross-species conservation and 3D mapping) can be used for the subsequent structural and functional annotations. Today, high-quality collections of protein isoforms are stored in UniProt, NCBI RefSeq, Ensembl databanks [11-13] and in more specific ones such as APPRIS, ISOexpresso and ASES [14-16].

Another important point is the existence of a single main protein isoform among several protein variants for each gene, which is called principal isoform or canonical protein. The canonical protein is identified by several criteria: experimental data on its functional role; data about its expression in different tissues of an organism; existence of the same combination of exons in orthologous proteins and in different curated databases. Although, in the annotated databases of proteomes [11-13] many canonical proteins are well distinguished from their isoforms, some of them are still poorly annotated.

Depending on the proteomes and quality of their annotation, the number of isoforms usually exceeds the amount of canonical proteins 2-3 times [11, 17]. At the same time, if to compare the number of proteins with the available experimental structural information, the situation is opposite. Almost all proteins in the Protein Data Bank [18] are canonical. Thus, due to a large gap between abundant sequence information and meager structural data on the isoforms, there is a growing interest in assessing the structural states and functional roles of alternative protein isoforms. As we have already mentioned, the sequence data on the isoforms are abundant. Therefore, if we want to get a global view of the structural-functional difference between the canonical proteins and their isoforms, apparently, the most appropriate approach is bioinformatics rather than the time-consuming experimental methods. In line with this need, during the last decade, significant progress has been achieved in the development of bioinformatics tools for large-scale structural and functional annotations of proteins. In the early days of structural bioinformatics, the foremost efforts of researchers were devoted to proteins with globular 3D structures. However, today, it is becoming clear that non-globular protein regions, having either intrinsically disordered conformations, membrane domains, elongated structures with tandem repeats or being aggregation-prone also have important functional roles [19-21]. Thus, an accurate structural and functional prediction of protein molecule can only be achieved when accounting for all these structural states. Recently, in line with this need, we developed a computational pipeline called TAPASS, which was designed to do just that [20]. The TAPASS pipeline is using known cutting-edge predictors able to detect intrinsically disordered regions (IDRs), transmembrane regions, signal peptides, conserved structured domains, short linear motifs (SLiMs) and aggregation-prone regions in protein sequences. The main novelty of this tool is a more precise prediction of aggregation-prone regions by taking into consideration the other known or predicted structural states. Moreover, the appearance of the AlphaFold program [22] opened up the possibility to model a large number of high-confidence structures of the isoforms. This artificial intelligence program, in a short time, became the gold standard computational technique for prediction of the 3D structure of proteins based on their sequence thanks to its accuracy competitive with experimental structures in a majority of cases.

In this study, by taking advantages of these state-of-the-art bioinformatics tools we systematically compared structural states of canonical proteins and isoforms. The analysis was performed on a large scale using 58 eukaryotic proteomes and provided a global view on the prevalence of each of these types of structures in canonical and isoform sets. Moreover, in some cases, our analysis proposed functional implications caused by

structural changes of the isoforms as well as the possibility of selective evolutionary pressure, to which they can be exposed for the functional roles.

2. Materials and Methods

2.1. Construction of datasets of canonical proteins and their isoforms

2.1.1. Main dataset

Construction of properly divided large datasets of canonical proteins and their isoforms represents a challenge because some proteins are still poorly annotated. To obtain large subsets of canonical proteins and their isoforms, we retrieved corresponding sequences from reference proteomes of 58 eukaryotic species (Supplementary materials S1) by using July 2020 release of UniProt databank [11]. Our choice was justified by the fact that UniProt contains a large combined set of several databases. The UniProt uses the following criteria to identify the canonical proteins: experimental data on their functional role; data about their expression in different tissues of an organism; existence of the same combination of exons in orthologous proteins and in different curated databases (https://www.uniprot.org/help/canonical_and_isoforms). First, we used option "Download all (FASTA (canonical & isoform)" to get 1 906 397 sequences including both canonical proteins and their isoforms. Second, we used "Download one protein sequence per gene" option to obtain a better-defined set of 1 244 044 canonical proteins. To avoid redundancy, we clustered the isoforms by CDhit [23] and removed the identical ones. This gave us 661,745 isoforms. Then we selected those isoform sequences, which had the same gene IDs as proteins from the canonical set and were highly similar BLAST(e-value < 10^{-35}) with them [24]. As a result, we obtained a dataset of 263 475 canonical proteins and 565 942 isoforms, which was used in our analysis (Supplementary materials S2).

2.1.2. Dataset of proteins from cancer-related genes with well-documented expression levels

Not all proteins from the UniProt databank have information about their expression level. Therefore, we built a smaller set of canonical proteins and corresponding isoforms of human cancer-related genes with well-documented expression levels in both 22 normal and cancer tissues. For this purpose, we used ISOexpresso database [15]. Our dataset contains 82 canonical and 166 isoform proteins, which were used for evaluation of the correlation between aggregation and expression level of proteins.

2.1.3. Datasets for estimation of the structural difference in isoforms by using AlphaFold modelling

To evaluate the structural changes caused by the differences in the sequences (hereafter referred as difference regions) of the corresponding canonical and isoform proteins we used pairs of proteins with the difference regions inside of well-conserved structured domains. For this purpose, we chose human proteins annotated in SwissProt [25] and having evidence of existence at protein level (PE=1). The conserved structural domains were detected by using HMM library of the CATH databank [26]. At the next step, we selected CATH domains that overlapped with the difference regions. A CATH domain found in a canonical protein may be shortened in the isoform so that the remaining domain is not able to fold. Therefore, we considered only isoforms where 1) canonical CATH domain is shorter than 200 aa and at least 70% of the domain remains in the isoform, or 2) canonical domain is longer than 200 aa and at least 50% of the domain remains in the isoform. For the modelling, we subsequently selected 168 canonical proteins with 223 corresponding isoforms where the difference regions were longer than 20 AA and located inside of the CATH domains. Finally, to select the most conserved and studied domains, we run the 168 canonical proteins by local BLASTP against PDB sequences from 7 species (*P.troglodytes, B.taurus, M.musculus, R.norvegicus, D.rerio,*

D.melanogaster, C.elegans) and kept only those having more than 10 hits with E-value < 10^{-6}. As a result, we obtained 53 canonical human proteins with 63 corresponding isoforms for the prediction by the AlphaFold program.

Subsequently, the 3D structures of the isoforms were predicted by AlphaFold Colab [27]. The structural models of the canonical proteins were obtained from the AlphaFold database (https://alphafold.com/download#proteomes-section). The obtained structural models were analysed by using PyMol [28].

2.2. Bioinformatics tools used to annotate structural states of proteins

To annotate structural states of proteins, we used TAPASS pipeline, which includes several prediction tools. Structured domains were predicted by using HMM libraries (e-value $<10^{-3}$) of CATH. Intrinsically disordered regions were detected by IUPred [29] and an in-house BISMM filter, which chooses hydrophilic regions greater than 75% and proline-rich regions more than 25%. Signal peptide and transmembrane regions were predicted with SignalP and TMHMM, respectively [30, 31]. The tool also predicts amyloidogenic regions (aggregation prone motifs) by ArchCandy2.0 [32], TANGO [33] and PASTA 2.0 [34] with their default parameters. We detected short linear motifs (SLiMs) of degradation (degrons) by using motifs collected in the Eukaryotic Linear Motif (ELM) resource [35].

2.3. Detection of structural changes in and around the difference regions

All types of the difference regions (insertion, deletion, non-identical and mixed) can cause structural changes not only in the place of their location but also in the flanking regions with identical sequences. Most of the methods used in the TAPASS for structural annotation of canonical and isoform proteins detected these changes automatically. However, cases when deletions truncated CATH domains required additional rules (see 2.1.3). Application of these rules in our analysis affected prediction of structured/unstructured regions and Exposed Aggregation-prone Regions (EARs).

2.4. Analysis of tandem repeats in canonical proteins and isoforms

Tandem repeat regions were identified by MetaRepeatFinder (MRF) (https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=15) [36] tool in five proteomes (*H.sapiens, M.musculus, D.melanogaster, D.rerio, A.thaliana*). From several tandem repeat finders of MRF, we chose Regex, T-REKS [37] and TRUST [38] which are specialized in the detection of TRs with units of less than 3 residues, less than 15 residues and more than 15 residues, respectively. As a result, the combination of these finders detects all types of tandem repeats. The overlap between "difference" region and TR region was counted if they had at least one common residue.

3. Results and Discussion

3.1. Identification, classification and distribution of difference regions

Difference in the sequences of canonical proteins and their isoforms is quite specific in comparison with the differences between orthologous/paralogous proteins. Frequently, the differences between the orthologues represent point mutations and (or) short indels spread over the proteins. While canonical proteins and their isoforms always have region(s) with identical sequences (corresponding to the same exons) and relatively long fragments where sequences can be completely different (Figure1). To detect the difference regions, we generated pairwise alignments of canonical-isoform proteins by using Clustal Omega [39] and treated it by our in-house script (Supplementary materials S3).

We classified the differences between the canonical-isoform pairs into 4 groups choosing as a starting point canonical sequences: insertion, deletion, non-identical and

mixed (Figure 1). The "non-identical" regions have different sequences of the same length. "Mixed" regions are those, which have both amino acid substitutions and indels in the difference region. Sometimes, these regions also include identical regions shorter than 10aa.

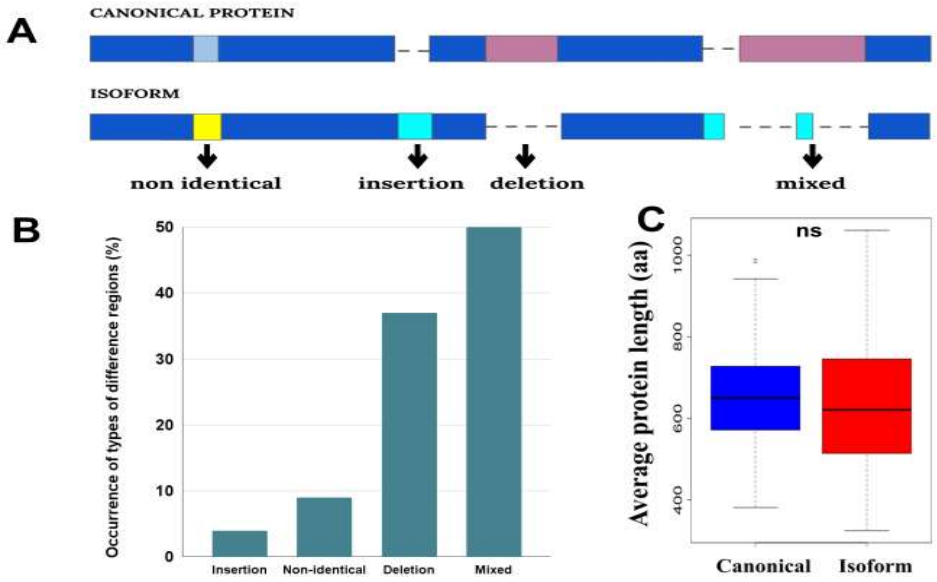

Figure 1. (A) Schematic representation of four groups of difference regions (dark blue and pink colors indicate identical and non-identical regions in the sequences, respectively). (B) Occurrence of types of the difference regions. (C) Distributions of the average length of canonical proteins and isoforms in proteomes. The distributions contain 58 points corresponding to the average length of each proteome. Here and elsewhere ns means non-significant difference with p-value > 0.05, *,**,*** and **** mean significant differences with p-value < 0.05, < 0.01, < 001 and < 0,0001, respectively.

The analysis showed that, the "mixed" difference region is the most common case followed by the deletions (Figure 1B). At the same time, a more detailed analysis of the "mixed" cases showed that it also contains a significant amount of deletions (68.6% of positions have deletions, 15.4% insertions and 16% amino acids). Because of the frequent deletions, in average, the isoforms are shorter in length than canonical proteins (Figure 1 C).

3.2. Distribution of structured and unstructured regions

Previous studies suggested that isoform proteins have a higher coverage of unstructured regions in comparison to the canonical proteins [40-42]. This conclusion suggested a lower level of involvement of isoforms in functional activity than of canonical ones. We examined this conclusion by using our datasets and TAPASS pipeline [20] (see 2.1.3). Our analysis showed that the proportion of proteins containing unstructured regions is slightly higher in the isoform set (Figure 2). The same tendency was observed when we compared coverage of unstructured regions in proteins. At the same time, both of these differences were not statistically significant. Thus, our results do not confirm the previous conclusions about higher number of unstructured residues in isoforms rather suggesting that the canonical proteins and their isoforms have the same ratio of residues in structured/unstructured states. This also suggests that during evolution isoforms preserve their structural domains, which are playing functional roles (Supplementary materials S4).

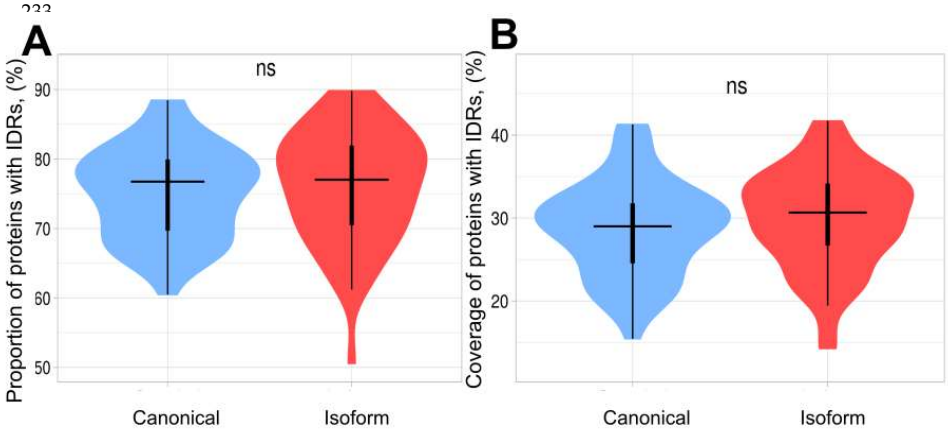

Figure 2. Violin plots of proportion and coverage of proteins containing IDRs in canonical and isoform proteins. The distributions contain 58 points corresponding to each proteome. (A) Proportion of proteins with IDRs in canonical proteins and isoforms. The difference between 2 sets is non-significant. (B) Coverage of IDRs in canonical proteins and isoforms, The coverage in isoforms is slightly higher, however, this difference is non-significant

3.3. Changes in subcellular localization

For the understanding of the functional role of a protein, it is important to know where it resides in the cell. There are a number of bioinformatics tools that can accurately predict the outcome of protein targeting in 4 major subcellular localizations: secreted proteins can be identified by SignalP [30], transmembrane regions (more exactly transmembrane helices) by TMHMM [31], nuclear proteins with nuclear localization signals can be found by regular expressions [35] and the remaining proteins as a rough approximation can be considered as cytosolic.

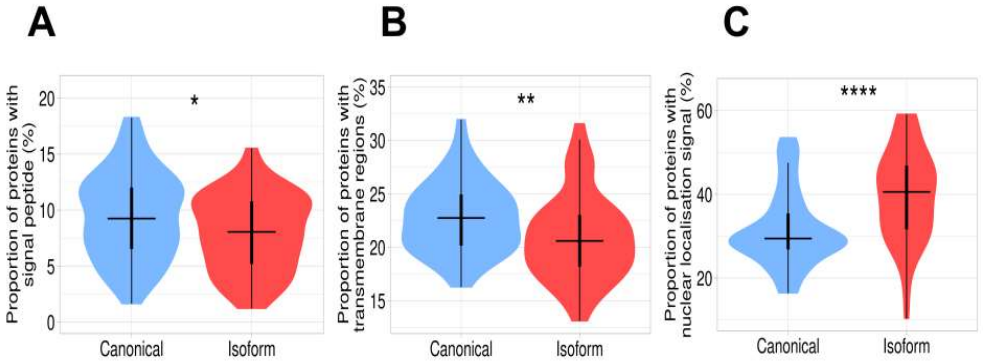

Figure 3. Difference in subcellular localization between canonical proteins and isoforms. (A) Proportion of proteins containing signal peptides. This value is significantly higher in canonical proteins than in isoforms. (B) Proportion of proteins containing transmembrane regions. The plot demonstrates a significant decrease of transmembrane proteins in the isoform set. (C) Proportion of proteins with nuclear localization signal. Isoforms have a remarkably high proportion of nuclear localization signals in comparison with canonical proteins.

Our analysis of the proportion of proteins with signal peptide showed that it is significantly lower in isoforms than in canonical proteins (Figure 3A). It suggests that in some cases the isoforms may maintain their globular functional domains but change their cellular localization from extracellular to cytosolic. A similar tendency was observed with the canonical proteins containing transmembrane helices (Figure 3B). Moreover, we found that the proportion of the nuclear localization signals in isoforms is significantly higher in

comparison with canonical proteins. It indicates that isoforms are more often localized in nucleus than canonical proteins (Figure 3C). The proportion of canonical proteins with transmembrane helices is higher than in isoforms, suggesting that a noticeable part of the isoforms loses their transmembrane localization. Parts of the difference regions that gain and loose signal peptides represent 2% and 4%, respectively. For the transmembrane helices, it is 2% and 7%. These changes may have important functional implications (Supplementary materials S4).

3.4. Proportion of aggregation-prone regions

Proteins are usually soluble and easily degraded by proteases after having performed their functions. However, some of them depending on the amino acid sequence and at certain condition can assemble into stable, protease-resistant aggregates. These aggregates are linked to serious diseases, which include, but are not limited to, Alzheimer's disease, Parkinson's disease, type II diabetes and rheumatoid arthritis [43]. Moreover, protein aggregation can be "functional" and play a central role in Liquid–liquid Phase Separation (LLPS), a process that leads to the formation of membrane-less organelles [44-45]. Several computational programs for prediction of protein aggregation have been developed [46]. The most realistic evaluation of the aggregation potential requires prediction of motifs located within unstructured regions and being aggregation-prone, which we call "Exposed Aggregation-prone Regions" (EARs) [20]. Here, we analyzed the EARs in canonical proteins and isoforms. Our interest in this analysis was also because, in general, canonical proteins have a higher level of cellular expression in comparison with their isoforms. It is reasonable to assume that to avoid aggregation, canonical proteins with the higher expression level may have the lower aggregation potential. The other reason of the higher aggregation potential of the isoforms may be truncation of native globular domains and unfolding of their remaining parts. For example, it was shown, that the p53 isoform, Δ133p53β, which is critical in promoting cancer activity is regulated through an aggregation-dependent mechanism [41]. The analyses of the truncated DNA-binding domain of Δ133p53β suggests that its remaining part most probably is unfolded and contains the EARs.

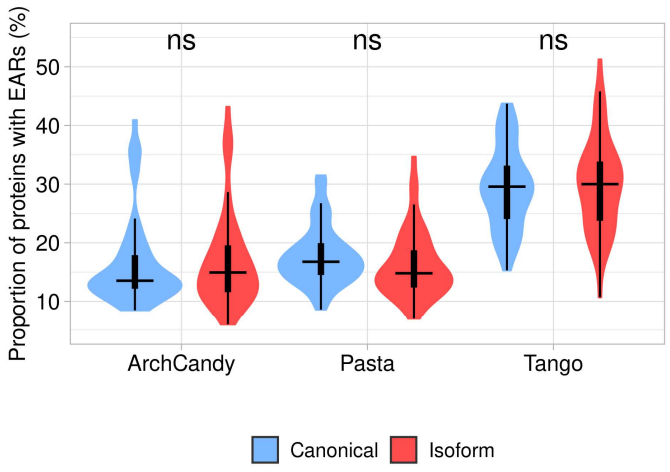

Figure 4. Proportion of EAR-containing proteins in canonical and isoform proteomes predicted by three tools (ArchCandy, Pasta, Tango). Differences between canonical proteins and isoforms are non-significant.

We estimated an average aggregation potential of canonical proteins and isoforms by proportion of EAR-containing proteins predicted by one of the predictors (ArchCandy, Pasta, Tango) in these two datasets. Our analysis revealed that the median value of

proportion for isoforms with EARs is almost the same as for canonical proteins (Figure 4). (Supplementary materials S4).

Although, it is accepted that the canonical proteins have higher expression levels than the isoforms [7,47], most proteins from our main dataset do not have reliable information about their expression level. Therefore, we analyzed also smaller sets with 82 canonical and 166 isoform proteins of human cancer genes with well-documented expression level in normal and cancer tissues (Supplementary materials S5, S6). These sets were used for evaluation of the correlation between aggregation and expression level of the proteins. The results confirm that average expression level of canonical proteins is significantly higher than of their isoforms. We also compared the relationship between expression level and aggregation potential of proteins in normal and cancer cells. The results of the analysis are shown in Figure 5. The expression of canonical proteins is higher in both normal and cancer cells. At the same time, expression level of all proteins slightly decreases in cancer cells. We also found that the proteins with EARs are expressed less in both normal and cancer cells than the ones without EARs. These results are in agreement with the assumption that to avoid aggregation, proteins with the higher expression level may have the lower aggregation potential.

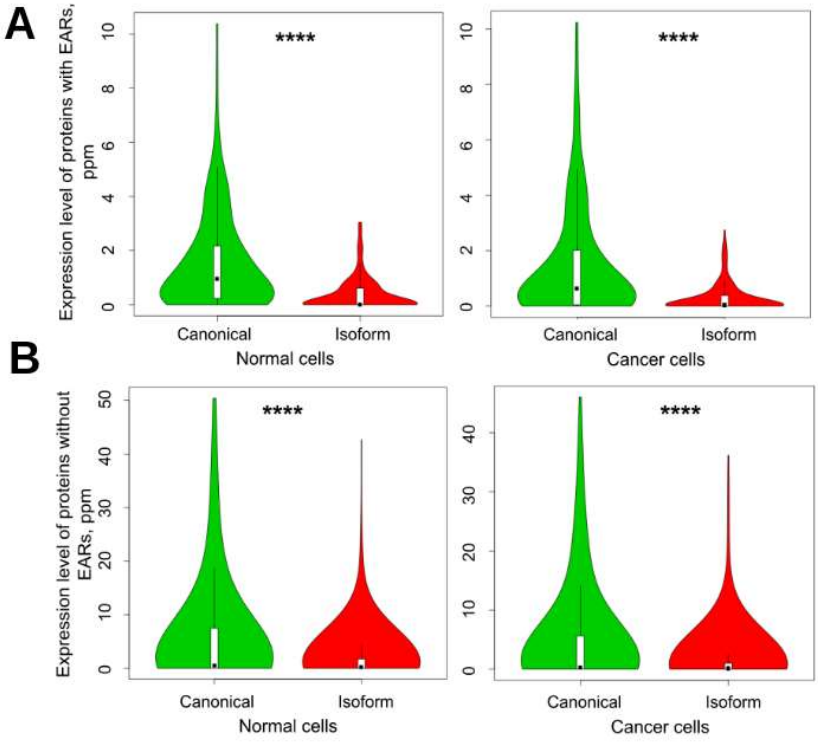

Figure 5. Violin plots of expression of canonical proteins and their isoforms in normal and cancer cells. (A) EAR-containing proteins and (B) non-EAR-containing proteins. EARs were predicted by using ArchCandy program. Mean levels of expression for EAR-containing canonical proteins and isoforms in normal cells were 1.565 and 0.386, respectively, and in cancer cells 1.490 and 0.306. For non-EAR-containing proteins, these values were 5.784, 1.773, and 4.984, 1.499 respectively. In accordance with T-test, all results were significant with p-values of less than 10^{-13}.

3.5. Canonical proteins have more degradation motifs than their isoforms

Abundance of proteins in the cell mostly depends on the balance of two opposite processes: expression and degradation. In general, canonical proteins have a higher level of cellular expression in comparison with their isoforms. It was interesting to understand if there is any difference between these proteins in terms of their degradation. The

experimental data on the protein degradation are limited and controversial. We compared canonical and isoform proteins *in silico* by analyzing the occurrence of degron motifs by TAPASS [20]. The degrons are short linear motifs that increase targeting of proteins for degradation [48-49]. We found that canonical proteins have higher proportion of degrons in comparison to the isoforms and this difference is statistically significant (Figure 6). (Supplementary materials S7)

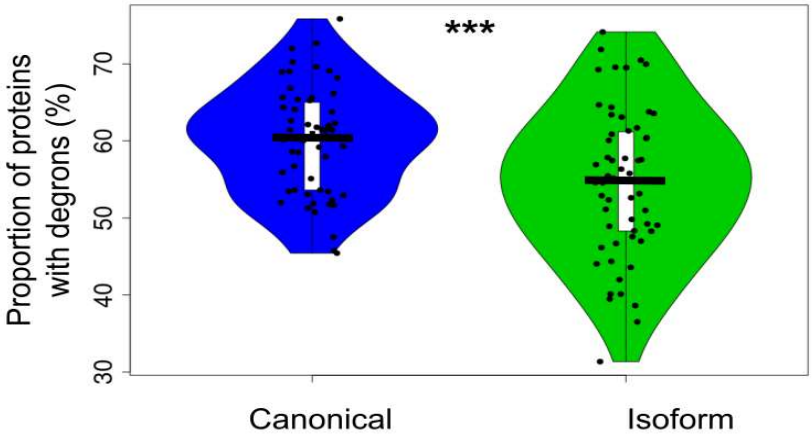

Figure 6. Proportion of canonical proteins and isoforms with degrons predicted by using SLiMs (T-test p-value = 0.00071). The distributions contain 58 points corresponding to each proteome. The proportion of degron-containing proteins is significantly higher in the canonical set than in the isoform one.

If more frequent occurrence of degrons in the canonical proteins causes their higher degradation rate in comparison with the isoforms, this may decrease the difference of the abundance between canonical proteins and isoforms. In its turn, similar level of the abundance may explain almost the same proportion of the aggregation-prone proteins predicted (Figure 4) for the canonical and isoform sets.

3.6. Occurrence of tandem repeats in canonical proteins and isoforms

Many protein sequences contain arrays of repeats that are adjacent to each other [50-51], so called tandem repeats (TRs). *S*everal authors have proposed that TRs might have evolved by exon duplication and rearrangement [52-53]. Therefore, it was interesting to get insight in the difference between canonical proteins and isoforms at these particular regions. We detected TRs in five well-annotated proteomes (*H.sapiens, M.musculus, D.melanogaster, D.rerio, A.thaliana*) by using MetaRepeatFinder (MRF) (https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=15). These proteomes contain in total 44357 canonical proteins. We found that a large part (43%) of them contains at least one TR region, and each TR-containing protein has, in average, about two TR regions. Comparison of the occurrence of the TR regions in canonical proteins and isoforms revealed that isoforms have less TR regions than canonical proteins (0.5 vs 0.81 TR region per protein) (Figure 7A). It is especially noticeable for TRs with a repeat length of 4-10 residues (Figure 7B). Partially, the decrease of TRs in the isoforms can be explained by the fact that among the differences between canonical proteins and isoforms we predominantly observed deletions (see section 3.1.). It was interesting to study the relationship between the location of the TRs and the difference regions. Our analysis showed that among the difference regions detected in the aligned pairs, a significant part (35%) overlaps with TRs.

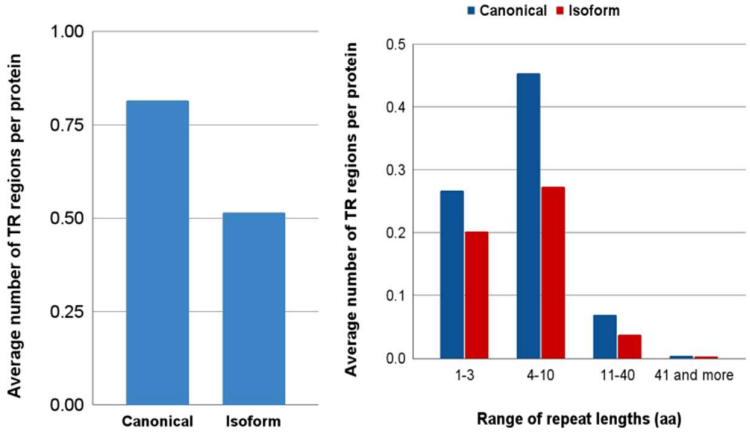

Figure 7. (A) Average number of tandem repeat regions determined per protein by MRF tool; (B) Distribution of proteins with tandem repeat by the length of their repetitive units.

3.7. Differences within the 3D structures of canonical proteins and isoforms predicted by AlphaFold.

Our proteome-wide analysis provides a global view on the canonical-isoform protein difference. At the same time, it is also interesting to investigate these changes from within the 3D structures down to the atomic details. In orthologous and paralogous proteins, the difference in the amino acid sequences of more than 30 % of identity may guarantee the same structural fold [54]. However, the character of the differences between canonical and isoform sequences is quite specific. They are identical at the location of the same exons, however, in the places of alternative splicing they can have completely different sequences. This "mosaic" arrangement may trigger significant structural and functional changes.

Given the fact that almost all proteins with experimentally determined 3D structures are canonical, the comparison requires molecular modelling of isoform structures. Previously, this type of modelling of the isoform structures and their comparison with the structures of the corresponding canonical proteins was described for some particular proteins [10]. Today, with the development of an artificial intelligence program AlphaFold [22], the scientific community got an opportunity to build high-quality structural models on a large-scale. Here, we applied AlphaFold program to obtain structural models of the isoform proteins. It was especially interesting to examine cases when the difference regions between the isoform and canonical proteins are conserved in several organisms and located within well-conserved structured domains. For the modelling, we used human proteins. To evaluate the cross-species conservation, we used 7 species from the Animal Kingdom (*P.troglodytes, B.taurus, M.musculus, R.norvegicus, D.rerio, D.melanogaster, C.elegans*). We considered that AlphaFold structural models are reliable when their level of the confidence (pLDDT) was higher than 70%, they did not have disallowed backbone conformations and the inside residues of the structure were predominantly apolar and did not have charged residues, which were not involved in the ionic bonds. The detection of unstructured regions was based on criteria used in TAPASS [20]. Several isoforms had difference regions outside of the well-conserved structured domains while inside of these domains they were identical between each other. Each group of these isoforms were reduced to one representative case. As a result, we compared the 3D structures of 50 canonical human proteins with 51 structural models of

the corresponding isoforms predicted by AlphaFold. This allowed us to classify the 3D structure transformations into four subgroups.

3.7.1. Exon deletions with the preservation of the overall structure

Proteins with tandem repeats

Though most of the selected proteins have globular structures, non-globular structures built of tandem repeats were found in 26% (13 of 51) of the cases. In the analyzed proteins with the difference regions inside of the complete structure, the most frequent situation is deletion of one repetitive unit. As a rule, these changes (also with any integer number of the repeats) does not cause serious structural perturbations (Figure 8A). These cases are observed in proteins with tandem repeats from Class III, IV and V [51,54-55]. In a few cases, the difference regions do not have an integer number of the repeats. This could lead to structural changes, if this difference is located in the middle of the repetitive structure. However, the isoform models showed that the change of the loop size between the repeats preserves the integrity of the whole structure (Supplementary materials S8, S9). In the other such cases, these difference regions are located at the terminal parts of the repetitive domains with no effect on the overall structure (Supplementary materials S8, S9). The described structural changes preserve the overall structure though create patches of new surfaces that can lead to modification of protein functions.

Globular proteins

Among 51 analyzed pairs, there are 20 globular structures, representing 38% of the cases, with the deletions of exons in the middle of the structure. In most of these cases, the deletion does not lead to critical structural transformations (Figure 8A). In some cases, it makes shorter loops preserving α-helices or β-strands; sometimes it removes one or several transmembrane helices. At the same time, these deletions can lead to changes in binding properties of the isoforms and (or) changes in the oligomerization states of the protein [56].

3.7.2. Exon substitutions that preserve the 3D structure

The other subgroup of four analyzed proteins (8% of the cases) is characterized by substitutions of exons. The size of the substituted exons is the same or almost the same and the sequences of canonical and isoform variants are not identical but similar. AlphaFold suggests that the new exons of the isoforms fits the native structure. This does not change the overall structure but leads to local changes on the molecular surface. This can be a basis for the modification of protein functions [57] (Figure 8B).

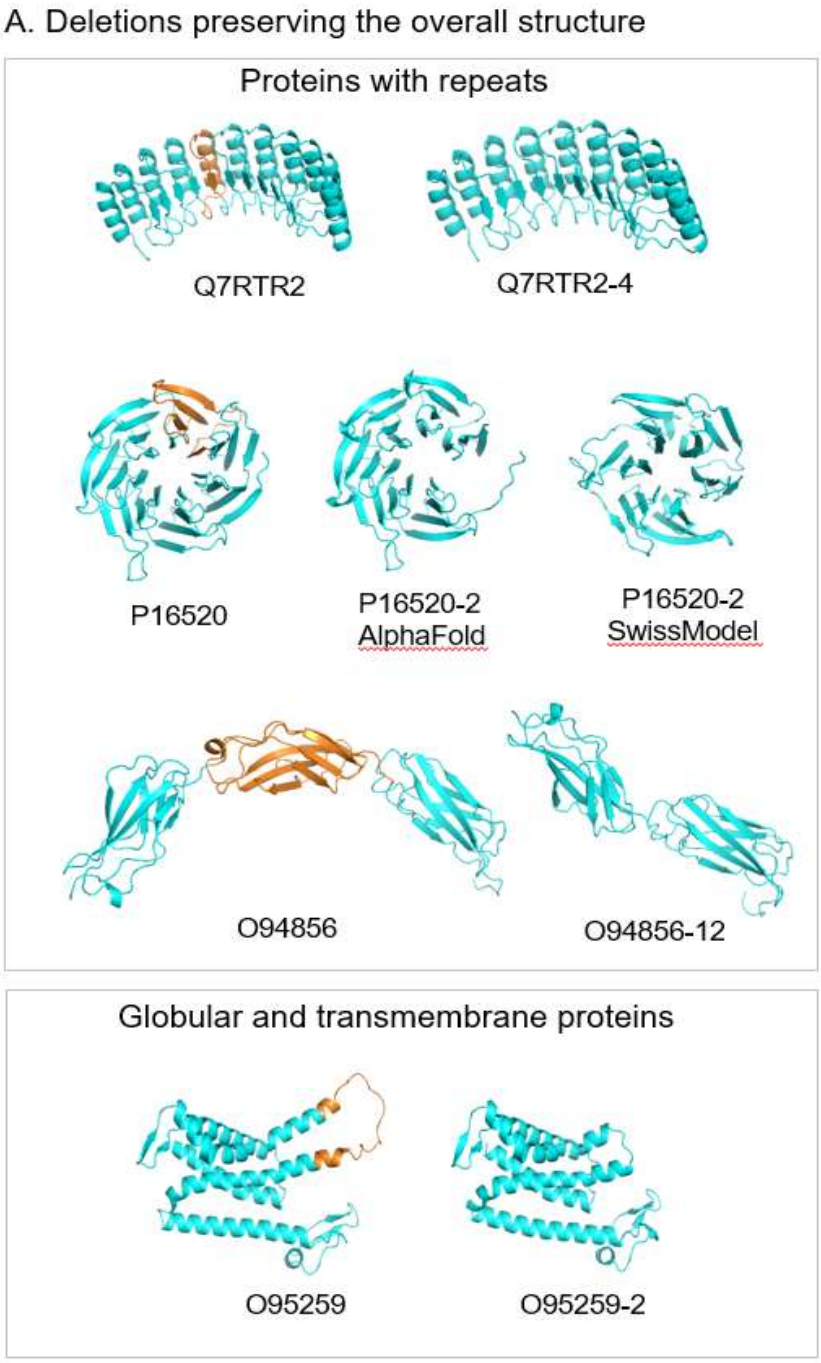

Figure 8. (continue)

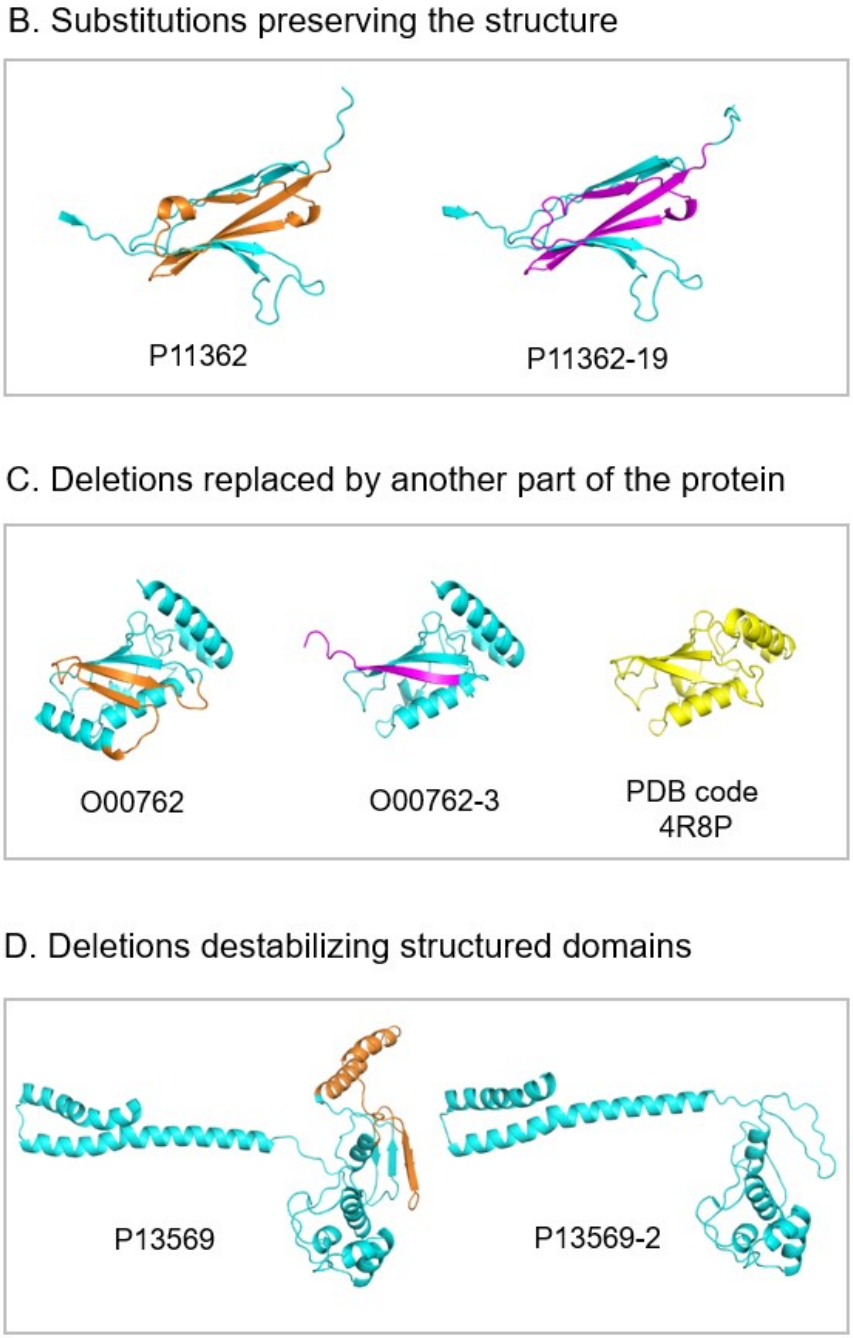

Figure 8. Ribbon representation of AlphaFold models of canonical proteins (left) and their isoforms (right). Fragments of canonical proteins deleted in the isoforms are in orange. Fragments of isoforms that substitute deleted fragments of the canonical proteins are in magenta. Representative structures of each subgroup from top to bottom are: Q7RTR2, LRR-protein of NLR family CARD domain-containing protein 3; P16520, 7-bladed beta-propeller of Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-3. AlphaFold model of isoform represents 6-bladed structure with an open beta-propeller, SwissModel structure made based on the known 6-bladed structure (PDB code 1E1A) has closed beta-propeller; O94856, Neurofascin; O95259, Potassium voltage-gated channel subfamily H member 1; P11362, Fibroblast growth factor receptor 1; O00762, Ubiquitin-conjugating enzyme E2 C, on the right, in yellow, the known crystal structure of ubiquitylation module similar to the truncated structure of the isoform in the center; P13569, Cystic fibrosis transmembrane conductance regulator.

3.7.3. Deletion that is substituted in the structure by another part of the molecule

We observed 6 of 51 cases (12%), where an exon deletion in the isoform removes a region that is critical for structural integrity of the globular domain. In the AlphaFold model of the isoform this part of the structure is filled by a new fragment, which, in the canonical protein belongs to an unstructured region. This suggests, that to provide structural diversity, proteins may have two or more neighboring regions, one is in the structure and another unstructured. If the first region is deleted in the isoform, the second one can dock into the structure, preserve it and modify the function. (Figure 8C)

3.7.4. Deletions that destabilize structured domains

We found eight cases (representing 16%) where exon deletions may destabilize the 3D structure of the isoforms. It mostly happened in the large multi-domain proteins. We assigned these examples to a separate subgroup. In these structures, the domain, which may be destabilized by the deletion of a critical part, can be transformed to an unfolded linker connecting the other globular domains. Instead, in the canonical structure these domains are connected by the structured domain (Figure 8D). In the case of canonical proteins with a single structured domain, the isoforms may represent intrinsically disordered proteins.

3.7.5. Limitations of AlphaFold in the interpretation of the conformational changes.

Our analysis revealed some limitations of AlphaFold modelling of the isoforms. For example, it is the case, when we try to distinguish between isoforms with exon deletions, which preserve the overall structure, from the ones that destabilize it. In most of the cases, we could not base our decisions on the confidence score pLDDT for the reason that even structures, which missed a large part of the domain, frequently had pLDDT score higher than 70%. These borderline cases were classified based on our visual analysis. In general, AlphaFold had tendency to build the isoform models that are very close to the canonical structures, but with missing parts corresponding to the deleted exons. One of these examples is shown in Figure 8A, where an isoform of the canonical 7-bladed beta-propeller of Guanine nucleotide-binding protein subunit beta-3 has 6 repetitive units. AlphaFold model of the isoform is almost identical to the canonical structure, but misses one blade leading to the structure with an open beta-propeller. However, SwissModel structure made based on the known 6-bladed structure (PDB code 1E1A) represents a closed 6-bladed beta-propeller. Such ambiguous cases cannot be resolved without experimental studies.

5. Conclusions

We took advantage of the progress achieved in the development of bioinformatics tools for large-scale structural annotations of proteins and examined the structural differences between canonical proteins and their isoforms. It became possible thanks to the TAPASS pipeline, which uses several state-of-the-art programs for prediction of structured domains, unstructured regions, transmembrane regions, aggregation-prone motifs [20]. Moreover, the availability of AlphaFold program [22] opened up the possibility to model a large number of the isoform structures. Altogether, our *in silico* analysis of 58 eukaryotic proteomes supported the concept that the majority of isoforms, similarly to the canonical proteins are under selective pressure for the functional roles. We also found that the proportions of proteins with signal peptide and transmembrane helices are lower in isoforms than in canonical proteins. This suggested that some isoforms lose their

transmembrane or extracellular localization and, eventually their functional roles. At the same time, we did not observe significant differences between canonical proteins and their isoforms in the occurrence of unstructured regions or aggregation-prone motifs. Our modelling of the isoform structures demonstrated that the AlphaFold program is perfectly suitable for investigations of the structural differences of splicing variants at atomic details. It was shown that frequently the isoform sequences being different from the canonical ones still can fold in similar structures. At the same time, the isoforms may have significant structural rearrangements, which can lead to changes of their functions. We suggested classification of the structural differences in the isoforms, which preserve the overall structure of the canonical proteins.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Conceptualization, A.V.K., Z.O. and I.S.; methodology, Z.O., A.V.K, T.F., and J.L.; software, Z.O., T.F and J.L; validation, T.S., G.A. and Z.O.; data curation, T.S., G.A. and Z.O.; writing—original draft preparation, Z.O. and A.V.K.; writing—review and editing, A.V.K., Z.O., G.A., T.F. and I.S..; supervision, A.V.K and I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by REFRACT project with Latin America in RISE program (2018-2023) H2020-MSCA-RISE-2018 to A.V.K, Azerbaijan National Academy of Sciences and The Ministry of Science and Education of Azerbaijan to Z.O.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thanks Prof. Layla Hirsh and Dr Nikola Arsic for discussion, careful reading of the manuscript and valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, E.T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **2008**, 456, 470–476. doi:10.1038/nature07509.
2. Pan, Q.; Shai, O.; Lee, L.J.; Frey, B.J.; Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **2008**, 40, 1413–1415.
3. Melamud, E.; Moult, J. Structural implication of splicing stochastics. *Nucleic Acids Research* **2009,** 37, 4862–4872. doi:10.1093/nar/gkp444.
4. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **2012**, 22, 1760–1774. doi: 10.1101/gr.135350.111.
5. Sánchez-Pla, A.; Reverter, F.; Ruiz de Villa, M.C.; Comabella, M. Transcriptomics: mRNA and alternative splicing. *Journal of Neuroimmunology* **2012**, 248, 23–31. doi: 10.1016/j.jneuroim.2012.04.008.
6. Uhlén, M.; Fagerberg, L.; Hallström, B.M.; Lindskog, C.; Oksvold, P.; et al. Tissue-based map of the human proteome. *Science* **2015**, 347, 1260419. doi: 10.1126/science.1260419.
7. Tress, M. L.; Abascal, F.; Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences* **2017**, 42, 98–110. doi: 10.1016/j.tibs.2016.08.008
8. Savosina, P.; Karasev, D.; Veselovsky, A.; Miroshnichenko, Y.; Sobolev, B. Functional and structural features of proteins associated with alternative splicing. *International Journal of Biological Macromolecules* **2020**, 147, 513–520. doi: 10.1016/j.ijbiomac.2019.09.241.
9. 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 Research* **2011**, 39, 1208–1219. doi:10.1093/nar/gkq843.
10. Birzele, F.; Csaba, G.; Zimmer, R. Alternative splicing and protein structure evolution. *Nucleic Acids Research* **2008**, 36, 550–558. doi:10.1093/nar/gkm1054

11. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **2021**, 49, D480–D489. doi: 10.1093/nar/gkaa1100.
12. O'Leary, N.A; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **2016**, 44, D733-745. doi: 10.1093/nar/gkv1189.
13. Cunningham, F.; Allen, J.E.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; et al. Ensembl 2022. *Nucleic Acids Res* **2022,** 50, D988–D995. doi.org/10.1093/nar/gkab1049.
14. Rodriguez, J.M.; Maietta, P.; Ezkurdia, I.; Pietrelli, A.; Wesselink, J.; et al. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research* **2013**, 41, D110–D117. doi: 10.1093/nar/gks1058.
15. Yang, I.S.; Son, H.; Kim, S.; Kim, S. ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. *BMC Genomics* **2016**, 17, 631. doi: 10.1186/s12864-016-2852-6.
16. Zea, D.J.; Richard, H.; Laine, E. ASES: visualizing evolutionary conservation of alternative splicing in proteins. *Bioinformatics* **2022**, 38, 2615–2616. doi.org/10.1093/bioinformatics/btac105.
17. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **2019**, 47, D506–D515. doi: 10.1093/nar/gky1049.
18. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; et al. The Protein Data Bank. *Nucleic Acids Research* **2000**, 28, 235–242. doi: 10.1093/nar/28.1.235.
19. Uversky V.N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Front Phys*. **2019**, 7:10. doi:10.3389/fphy.2019.00010
20. Falgarone, T.; Villain, É.; Guettaf, A.; Leclercq, J.; Kajava, A.V. TAPASS: Tool for annotation of protein amyloidogenicity in the context of other structural states. *J Struct Biol* **2022**, 214, 107840. doi.org/10.1016/j.jsb.2022.107840.
21. Uversky V.N. Typical Functions of IDPs and IDPRs. In *Intrinsically Disordered Proteins*, 1st ed.; Gomes, G.M.; Publisher: Springer Cham, 2014; pp. 13-33. https://doi.org/10.1007/978-3-319-08921-8.
22. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, 596, 583–589. doi.org/10.1038/s41586-021-03819-2.
23. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2021**, 28, 3150–3152. doi.org/10.1093/bioinformatics/bts565.
24. Boratyn, G.M.; Schäffer, A.A.; Agarwala, R.; Altschul, S.F.; Lipman, D.J. Domain enhanced lookup time accelerated BLAST. *Biol Direct* **2012**, 7, 12. doi: 10.1186/1745-6150-7-12.
25. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **2000**, 28, 45–48. doi: 10.1093/nar/28.1.45.
26. Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V.P.; Ashford, P.; et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res* **2021**, 49, D266–D273. doi: 10.1093/nar/gkaa1079.
27. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; et al. ColabFold: making protein folding accessible to all. *Nat Methods*. **2022**, 19(6):679-682. doi:10.1038/s41592-022-01488-1
28. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8 **2015.**
29. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research* **2018**, 46, W329–W337. doi: 10.1093/nar/gky384.
30. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **2011**, 8, 785–786. doi.org/10.1038/nmeth.1701.
31. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L.L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **2001**, 305, 567–580. doi.org/10.1006/jmbi.2000.4315.
32. Ahmed, A.B.; Znassi, N.; Château, M.T.; Kajava, A.V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimers Dement* **2015**, 11, 681–690. doi.org/10.1016/j.jalz.2014.06.007.
33. Rousseau, F.; Schymkowitz, J.; Serrano, L. Protein aggregation and amyloidosis: confusion of the kinds? *Curr Opin Struct Biol* **2006**, 16, 118–126. doi.org/10.1016/j.sbi.2006.01.011.
34. Walsh, I.; Seno, F.; Tosatto, S.C.E.; Trovato, A. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research* **2014**, 42, W301–W307. doi: 10.1093/nar/gku399.
35. Kumar, M.; Michael, S.; Alvarado-Valverde, J.; Mészáros, B.; Sámano-Sánchez, H.; et al. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res* **2022**, 50, D497–D508. doi.org/10.1093/nar/gkab975.
36. Richard, F.D.; Kajava, A.V. TRDistiller: a rapid filter for enrichment of sequence datasets with proteins containing tandem repeats. *J Struct Biol* **2014**, 186, 386–391. doi.org/10.1016/j.jsb.2014.03.013.
37. Szklarczyk, R.; Heringa, J. Tracking repeats using significance and transitivity. *Bioinformatics* **2004**, 20, i311–i317. doi:10.1093/bioinformatics/bth911.
38. Jorda, J.; Kajava, A.V. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **2009**, 25, 2632–2638. doi:10.1093/bioinformatics/btp482.
39. Madeira, F.; Pearce, M.; Tivey, A.R.N.; Basutkar, P.; Lee, J.; et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **2022**, gkac240 doi:10.1093/nar/gkac240.
40. Colak, R.; Kim, T.; Michaut, M.; Sun, M.; Irimia, M.; et al. Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. *PLoS Comput Biol* **2013**, 9, e1003030. doi:10.1371/journal.pcbi.1003030.

41. Arsic, N.; Slatter, T.; Gadea, G.; Villian, E.; Fournet, A.; et al. Δ133p53β isoform pro-invasive activity is regulated through an aggregation-dependent mechanism in cancer cells. *Nat Commun* **2021**, 12, 5463. doi.org/10.1038/s41467-021-25550-2.
42. Uversky, V.N.; Dunker, A.K. Understanding protein non-folding. *Biochim Biophys Acta.* **2010**, 1804(6):1231-64. doi:10.1016/j.bbapap.2010.01.017.
43. Pepys, M.B. Amyloidosis. *Annu Rev Med*. **2006,** 57(1):223-241. doi:10.1146/annurev.med.57.121304.131243.
44. Tsang, B.; Pritišanac, I.; Scherer, S.W.; Moses A.M.; Forman-Kay, J.D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell*. **2020**, 183(7):1742-1756. doi:10.1016/j.cell.2020.11.050.
45. Uversky, V.N.; Protein intrinsic disorder-based liquid-liquid phase transitions in biological systems: Complex coacervates and membrane-less organelles. *Adv Colloid Interface Sci.* **2017,** 239:97-114. doi: 10.1016/j.cis.2016.05.012.
46. Kotulska, M.; Wojciechowski J.W. Bioinformatics Methods in Predicting Amyloid Propensity of Peptides and Proteins, In *Computer Simulations of Aggregation of Proteins and Peptides,* 1st ed.; Li, M.S., Kloczkowski, A., Cieplak, M., Kouza, M., Publisher: Methods in Molecular Biology,Humana,New York, NY USA, 2022; Volume 2340, pp 1-15. doi.org/10.1007/978-1-0716-1546-1_1.
47. Ezkurdia, I.; Rodriguez, J.M.; Carrillo-de Santa Pau, E.; Vázquez, J.; Valencia, A.; et al. Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res.* **2015**, 14(4): 1880–1887. doi:10.1021/pr501286b.
48. Ravid, T.; Hochstrasser, M. Diversity of degradation signals in the ubiquitin–proteasome system. *Nat Rev Mol Cell Biol*. **2008**, 9(9):679-689. doi:10.1038/nrm2468.
49. Varshavsky, A. N-degron and C-degron pathways of protein degradation. *Proc Natl Acad Sci USA*. **2019,** 16(2):358-366. doi:10.1073/pnas.1816596116.
50. Andrade, M.A.; Perez-Iratxeta, C.; Ponting, C.P. Protein repeats: structures, functions, and evolution. *J Struct Biol*. **2001,** 134(2-3):117-131. doi:10.1006/jsbi.2001.4392.
51. Kajava, A.V. Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*. **2012,** 179(3):279-288. doi:10.1016/j.jsb.2011.08.009.
52. Paladin, L.; Necci, M.; Piovesan, D.; Mier, P.; Andrade-Navarro, M.A.; et al. A novel approach to investigate the evolution of structured tandem repeat protein families by exon duplication. *J Struct Biol*. **2020**, 212(2):107608. doi:10.1016/j.jsb.2020.107608
53. Liu, M.; Grigoriev, A. Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet*. **2004**, 20(9):399-403. doi:10.1016/j.tig.2004.06.013
54. Lesk, A.M.; Levitt, M.; Chothia, C. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng Des Sel* **1986,** 1, 77–78. https://doi.org/10.1093/protein/1.1.77.
55. Paladin, L.; Bevilacqua, M.; Errigo, S.; Piovesan, D.; Mičetić, I.; et al. RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Research*. **2021**, 49(D1):D452-D457. doi:10.1093/nar/gkaa1097
56. Wise, H. The roles played by highly truncated splice variants of G protein-coupled receptors. *J Mol Signal*. **2012**, 7(1):13. doi:10.1186/1750-2187-7-13
57. Dardenne, E.; Pierredon, S.; Driouch, K.; Gratadou, L.; Lacroix-Triki, M.; et al. Splicing switch of an epigenetic regulator by RNA helicases promotes tumor-cell invasiveness. *Nat Struct Mol Biol.* **2012**, 19(11):1139–1146. doi:10.1038/nsmb.2390