Report on the Impact and Bioinformatics of Alternative Splicing

Samuel J. Koch
Problem Based Learning Bioinformatics (PBL)

June 26, 2025

1 Introduction

Alternative splicing (AS) is a fundamental biological process that significantly contributes to genome complexity and proteomic diversity in eukaryotic organisms. This process allows a single gene to produce multiple messenger RNA (mRNA) and protein isoforms by selectively combining different exons into the mature mRNA. During splicing, different combinations of exons can be included or excluded, creating diverse mRNA molecules from the same gene. Each unique mRNA then serves as a template for producing distinct protein variants, called isoforms, which may have different functions, cellular locations, or regulatory properties. This mechanism is crucial for biological complexity, especially given that the number of protein-coding human genes is lower than initially predicted - alternative splicing provides a means to dramatically expand the functional diversity of the proteome from a relatively limited genetic repertoire. Defects in mRNA splicing patterns and their regulation have been implicated in various human diseases, including cancers.

This report aims to summarize key findings from two influential papers. Firstly, "Impact of Alternative Splicing on the Human Proteome" by Liu et al. (2017), which describes an integrative experimental approach to quantitatively assess how perturbations in mRNA splicing patterns alter the composition of the proteome. Secondly, "Bioinformatics of alternative splicing and its regulation" by Florea (2005), which provides an overview of bioinformatics methods developed to identify, characterize, and catalogue alternative

splicing and its regulatory elements.

2 Main Text

2.1 Types and Impact of Alternative Splicing Events

Alternative splicing is a process where different combinations of exons within a gene are spliced from the RNA precursor to be included in the mature mRNA, depending on factors such as tissue, developmental stage, or disease state. The resulting proteins can exhibit distinct, and sometimes antagonistic, functional and structural properties.

From a gene structure perspective, alternative splicing events can manifest in several ways, including:

- **Exon inclusion/exclusion**.
- **Alternative 5' exon end**.
- **Alternative 3' exon end**.
- **Intron retention**.
- **Alternative 5' and 3' untranslated regions (UTRs)**.

Liu et al. (2017) specifically investigated the impact of perturbed RNA splicing on the proteome by depleting **PRPF8**, a core spliceosome U5 small nuclear ribonucleoprotein (snRNP) component ([1], p. 1229). PRPF8 is essentially a key part of the spliceosome - the large molecular machine that actually performs the splicing of RNA. By reducing this critical component, the researchers could intentionally disrupt normal splicing patterns and then directly measure the consequences. This perturbation enabled a quantitative assessment of how splicing changes affect differential protein expression. Their findings highlight several key impacts on protein abundance:

- **Intron retention** is accompanied by **decreased protein abundance** ([1], p. 1237). Introns are the non-coding segments that are normally removed during RNA

processing. When introns are unexpectedly retained (kept in) the final mRNA, this consistently leads to less protein being made. This occurs because these faulty mRNAs might get stuck in the nucleus (unable to reach the protein-making machinery) or get destroyed by the cell's quality control system called nonsense-mediated decay (NMD), which cleans up potentially problematic messages.

- **Alterations in differential transcript usage (DTU)** and **differential gene expression (DGE)** lead to **protein abundance changes proportionate to transcript levels** ([1], p. 1238). This suggests that changes in mRNA abundance play a dominant role in determining dynamic changes in protein levels in a system with perturbed alternative splicing.

2.2 Methods for Studying and Identifying Alternative Splicing

2.2.1 Experimental Methods

Liu et al. (2017) developed an **integrative approach** combining RNA sequencing (RNA-seq) with mass spectrometry ([1], p. 1229). This clever strategy allowed them to measure both the messages (mRNA) and the final products (proteins) simultaneously, enabling them to connect the dots between splicing changes and their effects.

- **RNA-seq** was used to comprehensively report transcriptomic changes, including intron retention, differential transcript usage, and gene expression ([1], p. 1229). This technique provides a global picture of all the mRNA transcripts being produced.

- **SWATH-MS (sequential window acquisition of all theoretical spectra-mass spectrometry)**, a data-independent acquisition (DIA) method, was employed to capture an unbiased, quantitative snapshot of the impact of constitutive and alternative splicing events on the proteome ([1], p. 1230). This mass spectrometry technique is very good at identifying and, crucially, quantifying lots of different proteins in complex samples. SWATH-MS enabled the identification and quantification of 14,695 peptides uniquely mapping to 2,805 protein-encoding genes, showing high reproducibility.

- The effects at the protein level were **validated by targeted SRM (selective reaction monitoring)**, a more sensitive but lower-throughput mass spectrometric approach ([1], p. 1233). While SWATH-MS can measure thousands of peptides simultaneously, SRM focuses on a smaller number of specific targets with higher precision, making it ideal for confirming the key findings.

- A significant challenge was **integrating transcriptomic and proteomic datasets**, particularly regarding peptide assignment ([1], pp. 1231-1233). The problem was that many peptides (protein fragments detected by mass spectrometry) could come from multiple different transcript isoforms of the same gene, making it unclear which transcript was actually responsible for producing the detected protein. The study devised a clever strategy that used information from RNA-seq experiments to guide these assignments. They realized that not all transcripts are created equal - for each gene, they focused only on the **major transcript** (the most abundant isoform being produced). This approach provided much more usable information and better correlations compared to trying to assign peptides to all differently used transcripts regardless of their expression levels, which just added noise to the analysis.

2.2.2 Bioinformatics Approaches for Identifying Splice Variants

Florea (2005) surveys various bioinformatics methods used to identify and catalogue alternative splicing events ([2], p. 57). These methods often involve comparing expressed DNA or protein sequences of different isoforms to detect insertions or deletions.

- **Direct comparison of cDNA and protein sequences** can reveal differences ([2], p. 57).

- **Comparison of exon-intron structures** from cDNA- or protein-genomic spliced alignments can distinguish between different types of AS events and provide genomic context ([2], p. 57).

- **Microarray data** from Affymetrix chips with multiple probes per gene can be used to identify splice variations by detecting differential expression levels of probes when

an exon is alternatively spliced ([2], p. 60).

Key sequence data resources include ([2], p. 57):

- **dbEST**: Database for expressed sequence tags.
- **RefSeq**: NCBI's collection of curated full-length mRNA sequences.
- **Mammalian Gene Collection (MGC)**: NIH initiative to clone and sequence full-length open reading frames.
- **UniProt**: Universal Protein Knowledgebase for protein sequences.

Specialized **alignment tools** are used to align cDNA sequences to genomic sequences, such as EST_GENOME, Sim4, Spidey, GeneSeqer, Blat, ESTmapper, MGAlignIt, and GMAP ([2], pp. 57-58).

Two main bioinformatics strategies for annotating full-length alternatively spliced transcripts are ([2], pp. 58-60):

- **Gene indices**: These are gene- or transcript-oriented collections of EST and mRNA sequences grouped by sequence similarity ([2], pp. 58-59). Examples include UniGene, TIGR Gene Indices, and GeneNest. Challenges include over-clustering, under-clustering due to insufficient sampling, and high computational cost ([2], p. 59).
- **Genome-based clustering and assembly**: This approach clusters spliced alignments of cDNA and protein sequences at genomic loci ([2], pp. 59-60). A particularly important innovation is the **splice graph**, which provides a systematic way to model and analyze all the possible splicing patterns within a gene ([2], p. 60). The splice graph represents a gene as a directed acyclic graph where exons are vertices and introns are arcs connecting them. Different splice variants correspond to different paths through this graph from start to finish. While this approach enables systematic enumeration of all possible transcript candidates, a major limitation is that some of the computationally predicted combinations may represent artificial constructs without biological relevance. Methods like AIR annotation pipeline and

ECgene attempt to address this by scoring and prioritizing candidates that are most likely to be biologically relevant based on the strength of supporting evidence ([2], p. 60). This genome-based method helps resolve many issues like contamination and sequencing errors found in the earlier gene indices approach by using the genome sequence as a reliable reference.

2.3 Regulation of Alternative Splicing and Bioinformatics Methods for Its Study

2.3.1 Regulation Mechanisms

Pre-mRNA splicing is carried out by the spliceosome, which includes small nuclear ribonucleoproteins (snRNPs U1, U2, U4-U6) and accessory proteins ([2], p. 62). Beyond the well-characterized 5' (donor) and 3' (acceptor) splice sites, other RNA elements, known as **cis-regulatory elements**, located in exons and introns, play a role in exon inclusion or exclusion ([2], p. 62). These elements include:

- **Exonic splicing enhancers (ESEs)**.
- **Intronic splicing enhancers (ISE)**.
- **Exonic splicing silencers (ESS)**.
- **Intronic splicing silencers (ISS)**.

These elements promote or repress splicing through the activity of bound regulatory proteins, often acting in competition with each other to determine whether an exon gets included or excluded ([2], p. 62).

- **Splicing activators**, such as **SR (Ser/Arg) factors**, are proteins that promote exon inclusion. They bind to enhancer elements within the exon and help recruit other essential splicing machinery components like U2AF, essentially encouraging the spliceosome to include that exon in the final mRNA ([2], p. 62).

- **Splicing repressors**, largely from **hnRNP protein families** (e.g., hnRNP I/PTB, hnRNP A/B, NOVA-1), work in the opposite direction to promote exon exclusion. They can interfere with spliceosome assembly, block the action of nearby enhancers, or physically loop the exon to make it inaccessible to the splicing machinery ([2], pp. 62-63). The final outcome depends on the balance between these competing activating and repressing forces.

2.3.2 Bioinformatics Methods for Identifying Regulatory Elements

Methods for identifying splicing regulatory elements are still developing ([2], pp. 63-65). These include:

- **Consensus sequences**: Early work targeted purine-rich exonic elements, but these simple representations can only specify which nucleotides are allowed at each position and cannot quantify how much one nucleotide is preferred over another ([2], p. 63).

- **Position Weight Matrices (PWMs)**: A more sophisticated model that captures the relative preferences for each nucleotide at every position within a regulatory motif. PWMs are constructed from the frequency of each nucleotide at each position across a collection of experimentally identified binding sites ([2], p. 64). They provide a scoring system where higher scores indicate stronger matches to the motif. For example, PWMs for important splicing factors like SF2/ASF, SC35, SRp40, and SRp55 have been experimentally derived.

- **Statistical over-representation of k-mers**: This computational approach searches for short DNA sequences (k-mers, typically 6-8 nucleotides long) that appear more frequently in certain contexts than would be expected by chance. For instance, researchers compare the frequency of sequences in exons versus introns, or in weakly spliced exons versus strongly spliced ones, to identify potential regulatory elements ([2], p. 64).

- **Motif discovery methods**: Algorithms like MEME and Gibbs sampling that can

find common sequence patterns in a set of related sequences without prior knowledge of what to look for. While useful, these methods may have limited predictive power when applied to large, diverse sequence sets with potentially weak regulatory signals ([2], p. 65).

- **Comparative studies**: A powerful approach that identifies functionally important sequences by looking for regions that have been conserved during evolution. By comparing orthologous sequences between species like human and mouse, researchers can identify regulatory elements that have been maintained because they serve important functions ([2], p. 65). Studies have shown that intronic sequences flanking alternatively spliced exons are more strongly conserved than those around constitutive exons, suggesting they contain important regulatory information.

2.3.3 Functional Impact

Liu et al. (2017) provided a concrete example of the biological impact of functional mRNA isoforms through proteome diversity, focusing on a **switch event** where the identity of the major transcript changes across conditions ([1], p. 1235).

- They studied **lamin-associated polypeptide (LAP2)**, whose isoforms have distinct cellular locations and functions ([1], p. 1235). This protein serves as an excellent example because its different isoforms have been well-characterized and do very different jobs in the cell.

- After PRPF8 depletion, the dominant LAP2 isoform switched from LAP2b to LAP2a ([1], p. 1235). Essentially, the cell switched from making mostly the "b" version to mostly the "a" version. This change was observed at both mRNA and protein levels, with LAP2b decreasing and LAP2a increasing.

- Functionally, **LAP2b sits at the nuclear lamina** (the inner edge of the nucleus) and acts as a repressor, keeping certain genes quiet, including important targets of p53 and NF-κB - key regulators of cell growth, stress response, and inflammation. In contrast, **LAP2a is found throughout the nuclear interior** and is involved

in the structural organization of the nucleus, doing a different job in a different location ([1], p. 1235).

- The isoform switch led to altered LAP2 localization (less at nuclear lamina, more in nuclear interior) and **de-repression of p53 and NF-κB transcriptional targets** ([1], p. 1237). In simple terms, since LAP2b (the repressor) was reduced and LAP2a took over, the brakes came off those target genes, which became more active. This demonstrates the clear functional consequence of quantitative modulation of protein isoforms by alternative splicing.

2.4 Insights and Future Directions

The collective evidence from these papers highlights that alternative splicing is a critical determinant of genome complexity and plays a significant role in engendering proteomic diversity. The quantitative changes in protein abundance are strongly influenced by alterations in differential transcript usage and differential gene expression, often proportional to transcript levels. Intron retention, a specific form of alternative splicing, is consistently associated with decreased protein abundance.

Liu et al. (2017) demonstrated that usable information can be obtained from peptides mapping to multiple transcripts in the same gene, provided that transcript abundance information is considered ([1], p. 1233). The integrative methods developed for combining RNA-seq and quantitative mass spectrometry data provide a strong **foundation for future studies to examine the proteome-wide effects of altered RNA splicing associated with human diseases** ([1], p. 1238).

Looking forward, the ability to precisely measure how alternative splicing functionally tunes the human proteome opens exciting possibilities for therapeutic intervention and personalized medicine. Given that many diseases, including cancers and neurological disorders, are linked to splicing defects, several promising avenues emerge:

- **Therapeutic manipulation of splicing**: The potential to design therapies that correct faulty splicing patterns or selectively promote the production of beneficial isoforms over harmful ones.

- **Diagnostic tool development**: Creating new diagnostic approaches based on identifying specific isoform signatures that characterize disease states or predict treatment responses.

- **Personalized medicine applications**: Targeting specific splicing events that drive disease in individual patients, moving beyond one-size-fits-all treatments to precision therapies tailored to a patient's unique splicing profile.

- **Fine-tuning the proteome for health**: Developing strategies to therapeutically modulate the cellular protein landscape by controlling alternative splicing patterns, potentially correcting disease-associated protein imbalances.

The methodological advances demonstrated by Liu et al. (2017) represent a crucial step toward realizing these therapeutic possibilities by providing the quantitative framework needed to understand and eventually manipulate the relationship between splicing patterns and proteome composition.

3 Summary

Alternative splicing is a central regulatory mechanism that significantly expands the diversity of proteins encoded by the human genome, functionally tuning the proteome. The work by Liu et al. (2017) represents a major breakthrough in quantitatively demonstrating that changes in isoform usage directly manifest at the protein level, with outcomes ranging from proportional protein abundance changes with differential transcript usage and gene expression, to decreased protein levels associated with intron retention. This research provided clear examples of how alternative splicing switch events can alter protein localization and function, directly impacting cellular processes such as the regulation of p53 and NF-κB pathways.

The integrative experimental approaches combining RNA-seq with mass spectrometry have solved the long-standing challenge of quantitatively linking transcriptomic diversity to proteomic complexity. By demonstrating that splicing perturbations produce measurable and functionally significant changes in protein abundance, this work validates

alternative splicing as a mechanism that solves the paradox of biological complexity arising from a relatively limited number of genes.

Most importantly, these findings establish a foundation for therapeutic intervention. The ability to precisely measure how splicing changes affect protein landscapes opens new possibilities for developing diagnostics based on isoform signatures and designing therapies that manipulate splicing patterns to treat diseases. As many human diseases, including cancers and neurological disorders, are linked to splicing defects, the potential for personalized medicine approaches targeting specific splicing events represents an exciting frontier in precision therapeutics. The methodological framework established by Liu et al. provides the quantitative tools needed to move from simply cataloguing splicing diversity to actively manipulating it for therapeutic benefit.

4 References

[1] Liu, Y., Gonzàlez-Porta, M., Santos, S., et al. (2017). Impact of Alternative Splicing on the Human Proteome. *Cell Reports*, 20, 1229–1241.

[2] Florea, L. (2005). Bioinformatics of alternative splicing and its regulation. *Briefings in Bioinformatics*, 7(1), 55-65.