UNIVERSITÉ DE
SHERBROOKE

Multiple Spliced Alignment : A Consistency Based Approach

AUTHOR

Chakirou ALABANI (Group H)

BIN702 - Autumn 2024
Nadia Tahiri

Contents

1 Project description

Alternative splicing is now recognized as a fundamental process in eukaryotic organisms, enabling a single gene to generate multiple distinct transcripts. This mechanism affects the majority of human genes (*[Harrow et al., 2006, Tress et al., 2007, Kim et al., 2008, Wang et al., 2008, Chen and Manley, 2009]*) and is widely pervasive, playing a crucial role in enhancing the diversity of the proteome. Recent genome-wide studies indicate that 40–60% of human genes have alternatively spliced forms (*[Modrek and Lee, 2002]*). This extensive occurrence underscores the need to investigate the evolutionary and conservation patterns of transcript sets, as these insights are essential for a deeper understanding of the mechanisms that influence gene evolution.

Understanding the evolution of sets of alternative transcripts is a challenging task and requires automated methods and tools to compare sets of alternative transcripts from homologous genes. Alternative transcripts from homologous genes have traditionally been compared using pairwise spliced alignments (PSpAs). A PSpA aligns either a spliced RNA sequence or its DNA equivalent, the coding DNA sequence (CDS), with an unspliced DNA sequence. This approach helps identify homologous or corresponding exons between sequences, providing crucial insights for genome annotation and gene prediction (*[Stanke et al., 2006, Dunne and Kelly, 2018]*). Various methods have been developed to tackle different versions of the PSpA problem, which involves finding the best PSpA between two sequences based on a specific optimization function (see *[Jammali et al., 2019]*). However, PSpA is limited to comparing only two sequences at a time, making it unsuitable for examining the evolution of alternative splicing. This restriction also renders it ineffective for analyzing large databases, where multiple sequence comparisons are essential.

A logical extension of pairwise spliced alignment (PSpA) for studying the evolution of alternative spliced RNA sets is multiple spliced alignment (MSpA). This approach aligns a collection of spliced RNA sequences with their corresponding unspliced genomic sequences, allowing for detailed analysis of splicing and exon structures within the gene sequences. Unlike traditional multiple sequence alignment (MSA), which focuses solely on sequence similarity, MSpA incorporates the splicing and exonic architectures of the input genes. Just as MSA has greatly advanced our understanding of sequence evolution, MSpA is anticipated to reveal new insights into the evolution of alternative splicing and the relationships among alternative spliced RNA sets.

The MSpA framework also has practical applications for genome annotation by facilitating the identification of exons homologous to those in well-characterized species, thereby aiding in the prediction of conserved isoforms in newly annotated genomes.

In the current state of art, there are a few methods available for coomputong MSpAs, with SplicedFamAlignMulti (SFAM) being a leading approach (*[Jammali et al., 2022]*). SFAM include three extensions : SFAM_mblock, SFAM_tcoffee_p, and SFAM_tcoffee_m, each providing tailored functionalities for different alignment scenarios. A summary of SFAM methods mechanisms is shown in Figure 1.

In this report, we present SplicedFamAlignMultiCBA (SFAM_CBA), a greedy heuristic method developed to merge all pairwise spliced alignments (PSpAs) of known coding sequences (CDSs) and

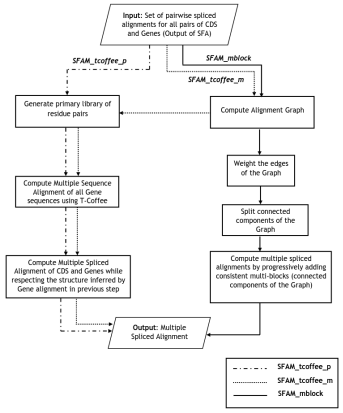

Figure 1: overview SFAM methods

gene sequences within a gene family into a multiple spliced alignment (MSpA).

2 Materials and methods

2.1 Characterization of the MSpA Problem

2.1.1 Pairwise Spliced Alignment (PSpA)

A Pairwise Spliced Alignment (PSpA) refers to the alignment of a Coding Sequence (CDS) with a gene sequence while accounting for the splicing structure. This alignment aims to identify homologous exon sequences and is formulated as a chain of blocks, where each block corresponds to pairwise alignments of segments from both the CDS and gene.

The significance of PSpA lies in its ability to highlight macroscopic alignments at the splicing level (exon intron structure) rather than at the nucleotide level, emphasizing differences in splicing trends and exon usage across gene families.

Definition: A PSpA consists of blocks, which are conserved segments where both the CDS and gene are included in the alignment. If a CDS is aligned with a gene in a block, it is termed a *conserved block*. Conversely, if only the CDS is included, it is referred to as a *deleted block*.

2.1.2 Multiple Spliced Alignment (MSpA)

An MSpA extends the concept of PSpA to include multiple CDSs and gene sequences, facilitating the identification of homologous exons across a broader set of sequences.

Definition: An MSpA of a set of CDSs C and genes G is represented as a chain of multiblocks $A = \{A[1], \ldots, A[n]\}$. Each multiblock $A[i]$ includes a key set key$(A[i])$, which is a subset of $C \cup G$,

mapping each sequence to its respective start and end locations (s_i^x, e_i^x).

The MSpA must satisfy several conditions:

- Each multiblock must contain at least one element.
- Segments from the same sequence must not overlap across multiblocks.
- The segments induced by the MSpA must fully cover the original CDSs.
- Segments from aligned genes must be consistent within the same multiblock.

2.1.3 Multiple spliced alignment problem

Input: A set of CDSs C; a set of genes G, a set of PSpAs $X = \{X_{c,g} \mid (c, g) \in C \times G\}$ for all pairs in $C \times G$.

Output: An MSpA A of C on G that maximizes the sum of the scores of induced PSpAs:

$$\sum_{(c,g)\in C\times G} S(A_{c,g})$$

(*Jammali et al. [2022]*).

Several methods are available for computing PSpAs between a gene and a CDS (*Jammali et al. [2019]*).

2.2 The SFAM-CBA algorithm

We now present our algorithm for constructing an MSpA for a set of CDSs C and a set of genes G, given a set of PSpAs $X = \{X_{c,g} \mid (c, g) \in C \times G\}$ for all pairs in $C \times G$ (the MSAPproblem). MSA methods commonly employ greedy heuristics, with the progressive alignment strategy (*Feng and Doolittle [1987]*) being one of the most widely used approaches. Distinctively, our method incorporates both the consistency approach and the triplet approach outlined in *CÃ©dric Notredame [2000]* for pairwise scoring. The consistency-based strategy evaluates how well the alignment of two residues or segments in a given pairwise alignment is supported by other precomputed pairwise alignments. Applying this consistency-based approach within a progressive alignment allows for the integration of information from all pairwise alignments at each step, helping to overcome the typical limitations associated with progressive alignment.

2.2.1 Boundaries refinement

Boundary refinement is a crucial step in the consistency based approch for MSpA, aiming to handle overlapping segment matches effectively. The refinement algorithm extends the principles established by *Halpern et al. (2002)*, ensuring that all parts of the original segment matches can be utilized.

As described in *Feng and Doolittle [1987]*, let $M = \{M_0, M_1, \ldots, M_{m-1}\}$ represent the set of segment matches, where each match $M_k = (S_i^{uv}, S_j^{xy})$ consists of segments S_i^{uv} from sequence S_i and S_j^{xy} from sequence S_j. The goal is to refine these segment matches into a set of submatches $M^* = \{M_0^*, M_1^*, \ldots, M_{m'-1}^*\}$ that cover the original matches.

In this context, the refinement process involves ensuring that the set of submatches M^* satisfies the conditions of tiling the original matches:

$$[u, v-1] = \bigcup_{M'_k \in M'_*} [u', v'-1] \quad \text{and} \quad [x, y-1] = \bigcup_{M'_k \in M'_*} [x', y'-1].$$

This ensures that each original match is fully covered by the refined submatches in M^*.

A resolved set of matches, denoted as R, is desired, where all segments are either disjoint or identical. In such a set, any $(S_i^{uv}, S_j^{xy}) \in R$ satisfies:

$$[u, v] \cap \text{supp}_S^i(R) = \{u, v\} \quad \text{and} \quad [x, y] \cap \text{supp}_S^j(R) = \{x, y\}.$$

The algorithm proceeds to process segment matches sequentially, building a node set V_i for each sequence S_i, initialized to the support set of the segments. By recursively identifying boundary positions and ensuring necessary cuts are made, the algorithm guarantees a minimum cardinality refinement without introducing superfluous cuts.

2.2.2 Computation of pairwise scores

The computation of pairwise scores of the set of PSpAs $X = \{X_{c,g} \mid (c, g) \in C \times G\}$ is vital for quantifying the similarity between aligned segments of sequences. This process relies on the refined segment matches obtained in the previous step.

Given a set of refined matches M^*, the pairwise score S_{ij} between sequences S_i and S_j can be calculated as follows:

$$S_{ij} = \sum_k \text{idty}(M_k^*),$$

where $\text{idty}(M_k^*)$ represents the identity score of the match M_k^* that covers the segment (S_i^{uv}, S_j^{xy}). The identity score can be derived from established scoring matrices adjusted for the context of the segments being compared.

To incorporate information from multiple sequences, a weight system can be applied, similar to the one discussed by Notredame et al. (1998). By considering intermediate sequences that support the alignment, the weights associated with each residue pair are aggregated, enhancing the reliability of the pairwise scores.

The overall weight W for a pair of aligned residues is computed by examining triplets of sequences and accumulating scores based on their alignments. The final weight reflects both the similarity of the sequences and their consistency across the dataset, facilitating a more robust alignment process.

2.2.3 MSpA Computation

The final stage involves the computation of the multiple sequence alignment (MSA) itself. This process is executed through a recursive algorithm that utilizes the refined pairwise scores obtained in the previous steps. The goal is to construct an alignment that maximizes the overall score while adhering to the constraints imposed by the segment matches.

Let T represent the phylogenetic tree that organizes the sequences based on their evolutionary relationships. The MSpA can be computed recursively as follows:

$$C(T_i) = \text{Align}(C(T_{\text{left}}), C(T_{\text{right}}), S),$$

where $C(T_i)$ is the column list of aligned sequences at node T_i and S is the pairwise score matrix derived from the previous computations.

Each internal node of the tree represents an alignment of its child nodes. The alignment function Align combines the column lists from the left and right child nodes, optimizing the alignment based on the accumulated pairwise scores:

$$\text{Align}(C(T_{\text{left}}), C(T_{\text{right}}), S) = \max_{\text{alignment}} \sum_{\text{pairs}} S_{ij},$$

where the sum is taken over all pairs of aligned segments, and the maximum is computed over all possible alignments.

This recursive approach allows the algorithm to build the MSpA incrementally, ensuring that the final alignment is optimal with respect to the defined scoring system. The use of a refined weight system ensures that the alignment reflects not only the local similarities but also the global relationships among all sequences, thereby enhancing the quality and reliability of the MSpA.

2.3 Experimental setup

A comprehensive and detailed pseudocode of each step in the algorithm is provided in the appendix. The implementation of the algorithm, along with the data used, is available on GitHub at

List of Figures

List of Tables

References

M. Chen and J. L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*, 10:741–754, 2009.

Jaap Heringa CÃ©dric Notredame, Desmond G. Higgins. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302:205–217, 2000. doi: 10.1006/jmbi.2000.4042.

M. P. Dunne and S. Kelly. OMGene: mutual improvement of gene models through optimisation of evolutionary conservation. *BMC Genomics*, 19:307, 2018.

D.-F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25:351–360, 1987.

J. Harrow, F. Denoeud, A. Frankish, et al. Gencode: producing a reference annotation for encode. *Genome Biology*, 7 Suppl 1:S4.1–S9, 2006.

S. Jammali et al. Splicedfamalign: Cds-to-gene spliced alignment and identification of transcript orthology groups. *BMC Bioinformatics*, 20:133, 2019.

Safa Jammali, AbigaÃ¯l Djossou, Wend-Yam D D OuÃ©draogo, Yannis Nevers, Ibrahim Chegrane, and AÃ¯da Ouangraoua. From pairwise to multiple spliced alignment. *Bioinformatics Advances*, 2 (1):vbab044, 2022. doi: 10.1093/bioadv/vbab044.

E. Kim, A. Goren, and G. Ast. Alternative splicing: current perspectives. *BioEssays*, 30:38–47, 2008.

B. Modrek and C. Lee. A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–19, Jan 2002. doi: 10.1038/ng0102-13.

M. Stanke et al. Augustus at egasp: using est, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, 7:S11, 2006.

M. L. Tress, P. L. Martelli, A. Frankish, et al. The implications of alternative splicing in the encode protein complement. *Proceedings of the National Academy of Sciences of the United States of America*, 104:5495–5500, 2007.

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

3 Appendix