Central Dogma and Splicing

exon1 intron1 exon2 intron2 exon3
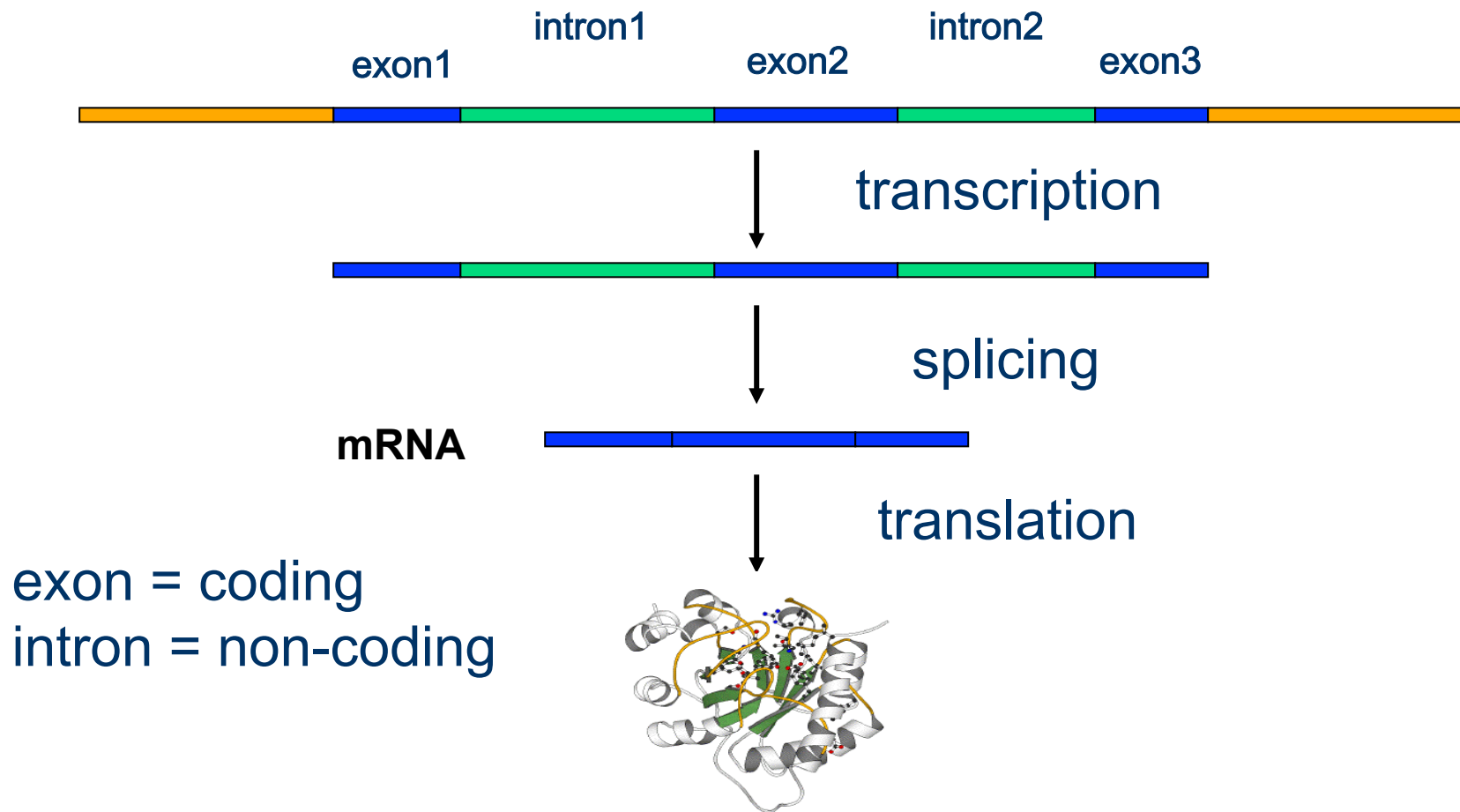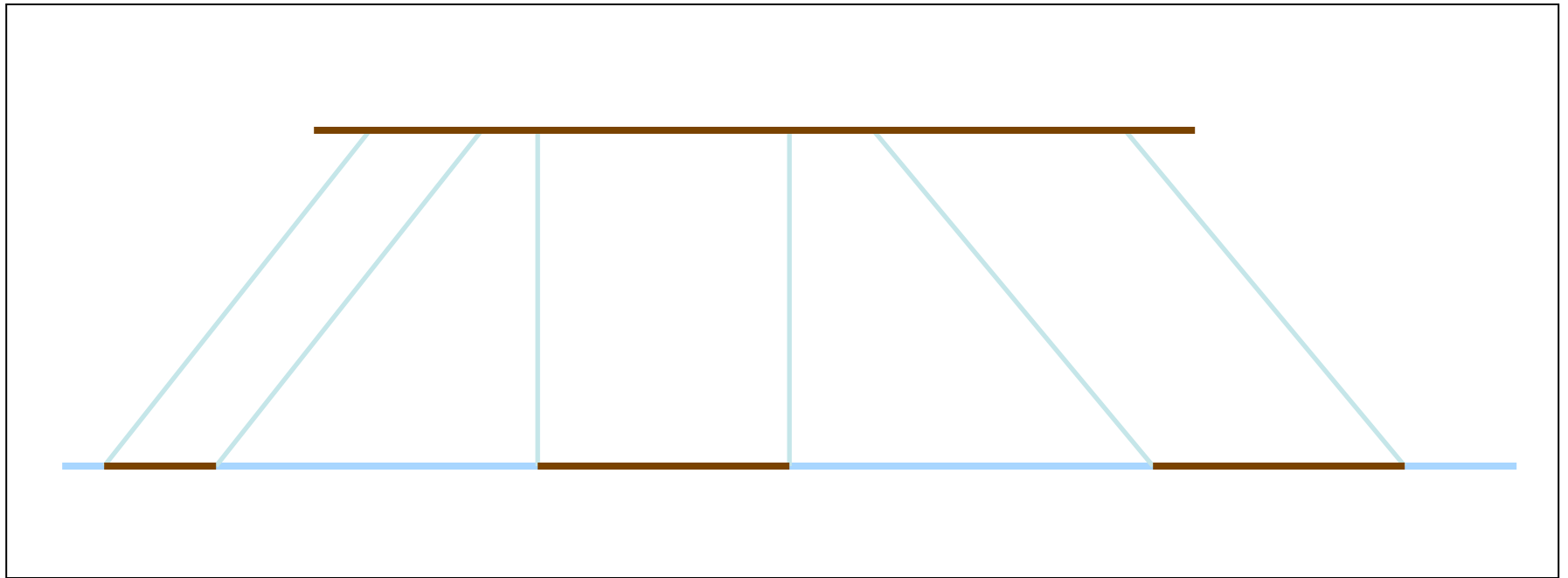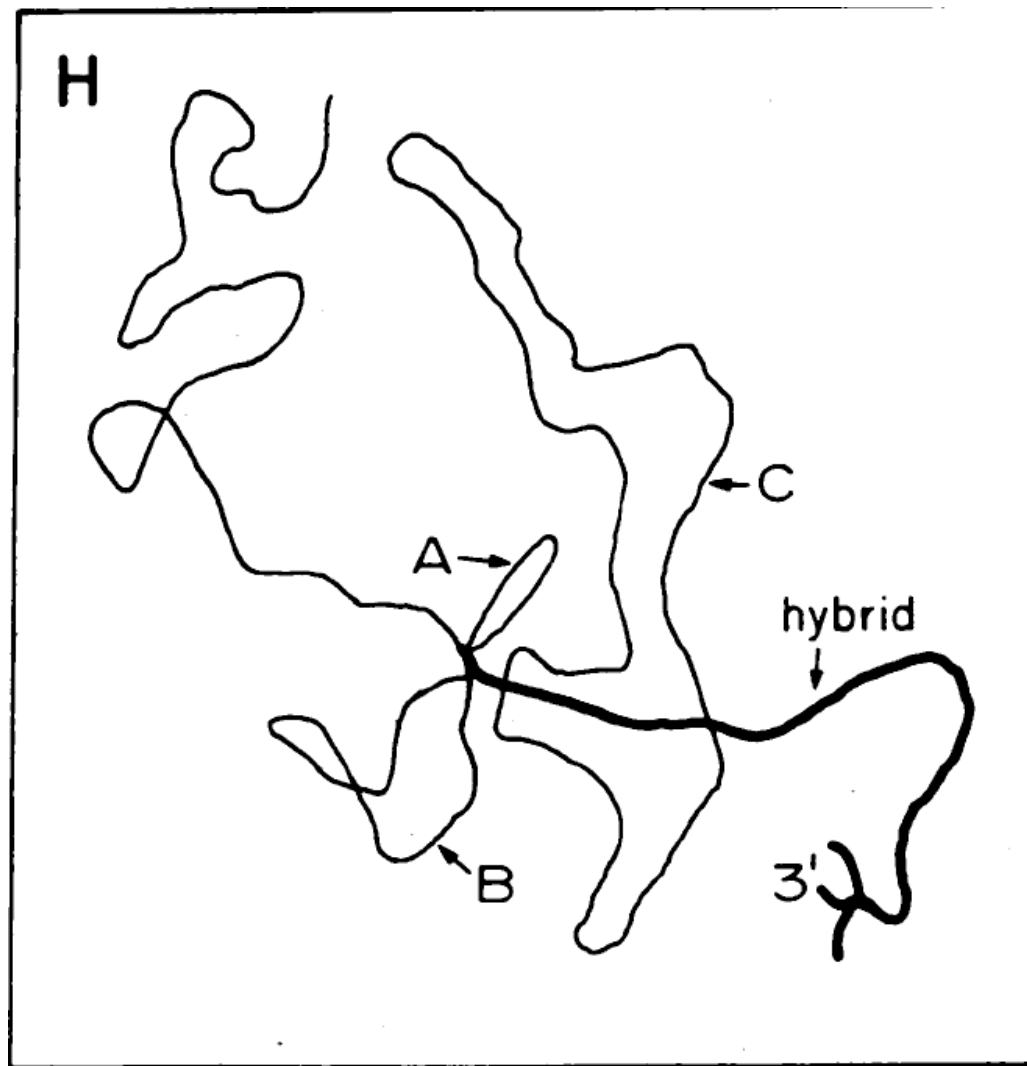
transcription

splicing

mRNA

translation

exon = coding
intron = non-coding

Comparing genes in two genomes

- Small islands of similarity corresponding to similarities between exons

Berget et al. (1977)

Gene finding

- **Problem:** Given a known mRNA sequence (e.g. in mouse) and an un-annotated genome sequence (e.g. in human), find a set of substrings of the genomic sequence whose concatenation best fits the mRNA (which is the homologous gene of the mouse gene in human).

Spliced sequence alignment
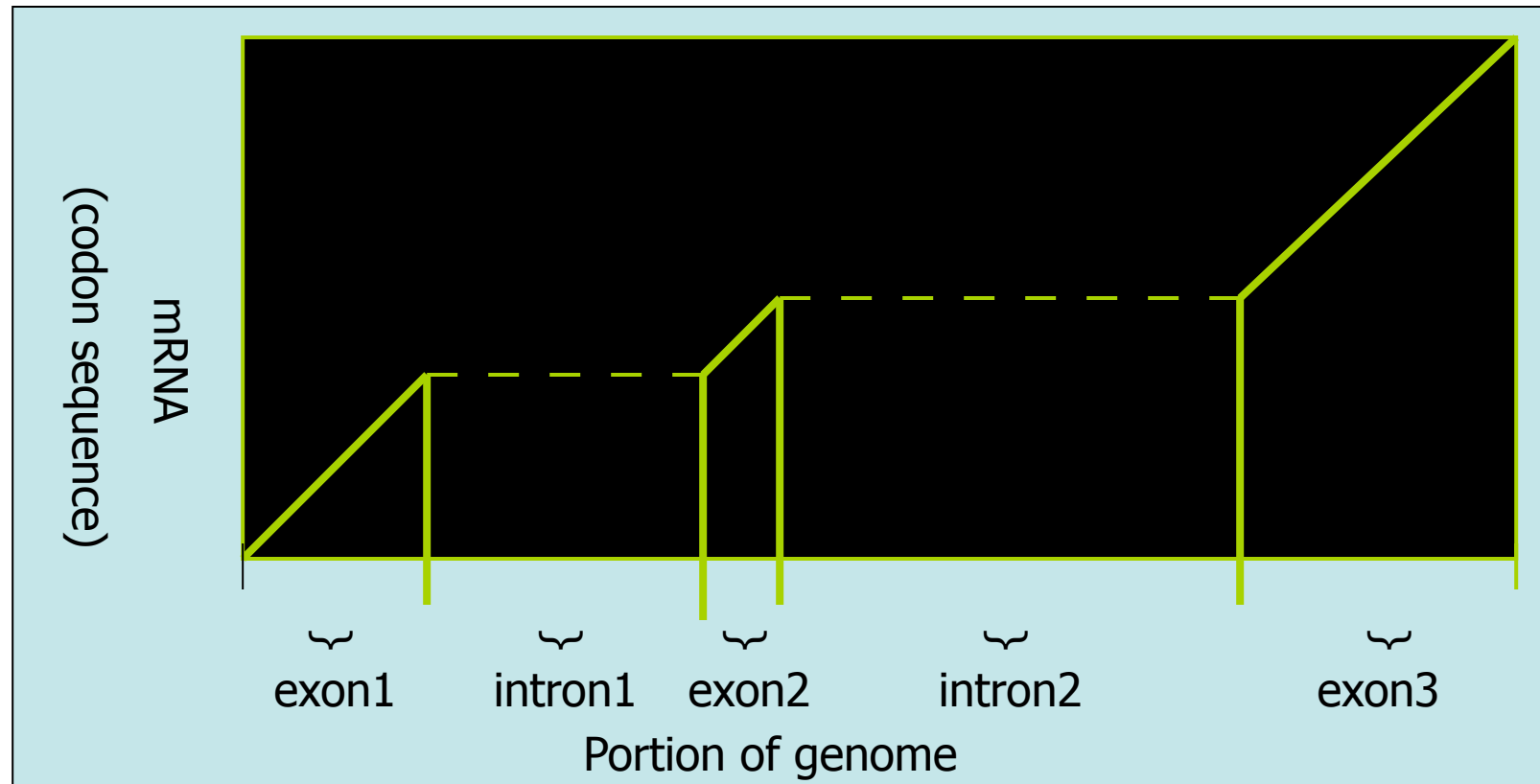
Spliced alignment problem: formulation

- **Goal**: Find a chain of blocks (putative exons) in a genomic sequence that best fits a target sequence

- **Input**: Genomic sequences G, target mRNA sequence T, and a set of putative exons B.

- **Output**: A chain of exons Γ such that the global alignment score between Γ^* and T is maximum among all chains of blocks from B.

  Γ^* - concatenation of all exons from chain Γ

"Splicing gap" edges to the alignment graph

Genome

mRNA
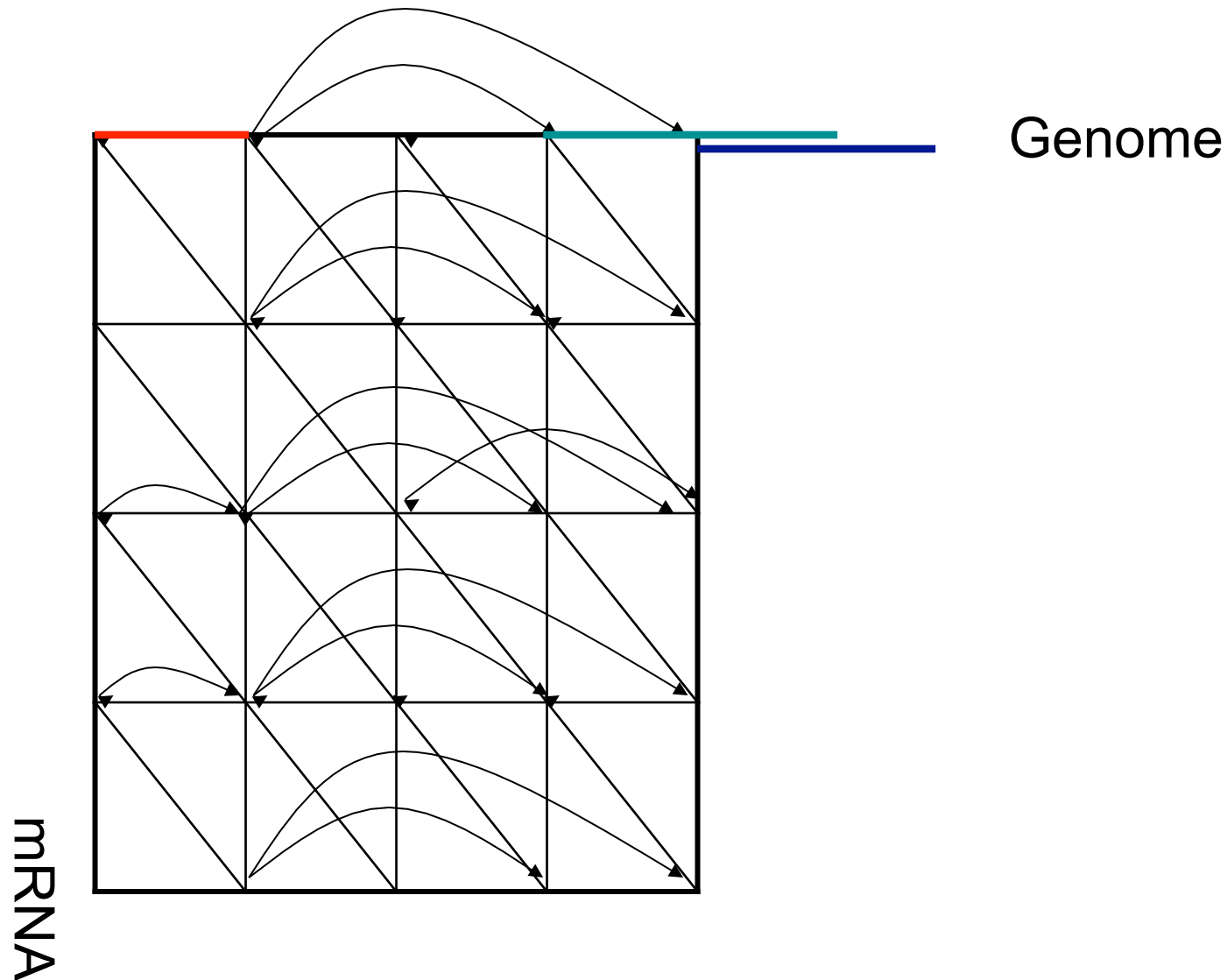
Spliced alignment recurrence

If *i* is not the starting vertex of block *B*:

- $S(i, j, B) =$

  $\max \{ \quad S(i - 1, j, B) - \textit{indel penalty}$

  $\qquad\quad S(i, j - 1, B) - \textit{indel penalty}$

  $\qquad\quad S(i - 1, j - 1, B) + \delta(g_i, t_j) \}$

If *i* is the starting vertex of block *B*:

- $S(i, j, B) =$

  $\max \{ \ S(i, j - 1, B) - \textit{indel penalty}$

  $\qquad \max_{\text{all blocks } B' \text{ preceding block } B} \ S(end(B'), j, B') - \textit{indel penalty}$

  $\qquad \max_{\text{all blocks } B' \text{ preceding block } B} \ S(end(B'), j - 1, B') + \delta(g_i, t_j)$

  $\qquad \}$

Spliced Alignment Solution

- After computing the three-dimensional table $S(i, j, B)$, the score of the optimal spliced alignment is:

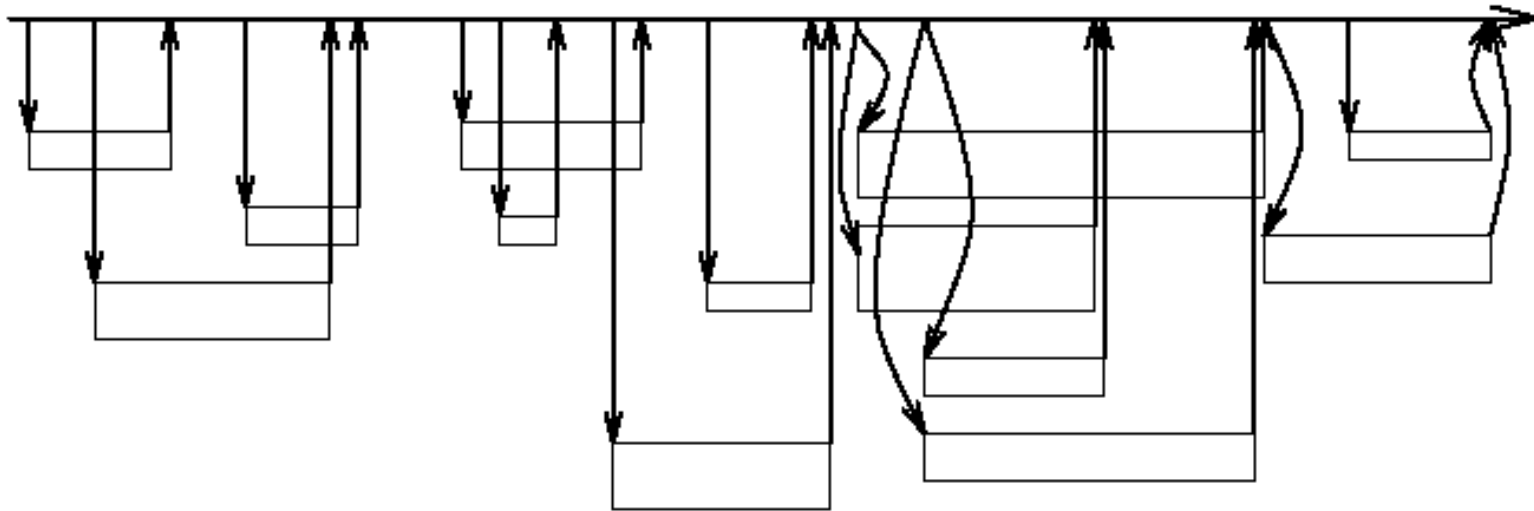$$\max_{\text{all blocks } B} S(end(B), length(T), B)$$

Spliced alignment: complexity

- Considering multiple *i*-prefixes leads to slow down. running time:

$$O(mn^2 |B|)$$

  where *m* is the target length, *n* is the genomic sequence length and |*B*| is the number of blocks.

Spliced alignment: a special case of network matching problem

- Input: a Directed Acyclic Graph (DAG, or network), G, with edges labeled by strings; and a string S.

- Output: a path in G, with the labeling sequence best fits S.

Spliced alignment of two genome sequences

- Input: two Directed Acyclic Graphs (DAG, or network), G_1 and G_2.

- Output: paths P_1 and P_2 in two graphs, respectively, with the labeling sequences best fits each other.

- Example: align entire human and mouse genomes while predicting genes in both sequences simultaneously as chains of aligned blocks (exons)

- This approach does not assume any annotation of either human or mouse genes.