3. Gene prediction

- All genes end on a stop codon
- A simple algorithm for gene prediction
- Searching for start and stop codons
- Predicting all the genes in a sequence
- Making the predictions more reliable
- Boyer-Moore algorithm
- Index and suffix trees
- Probabilistic methods
- Benchmarking the prediction methods
- **Gene prediction in eukaryotic genomes**

Eukaryotic genomes

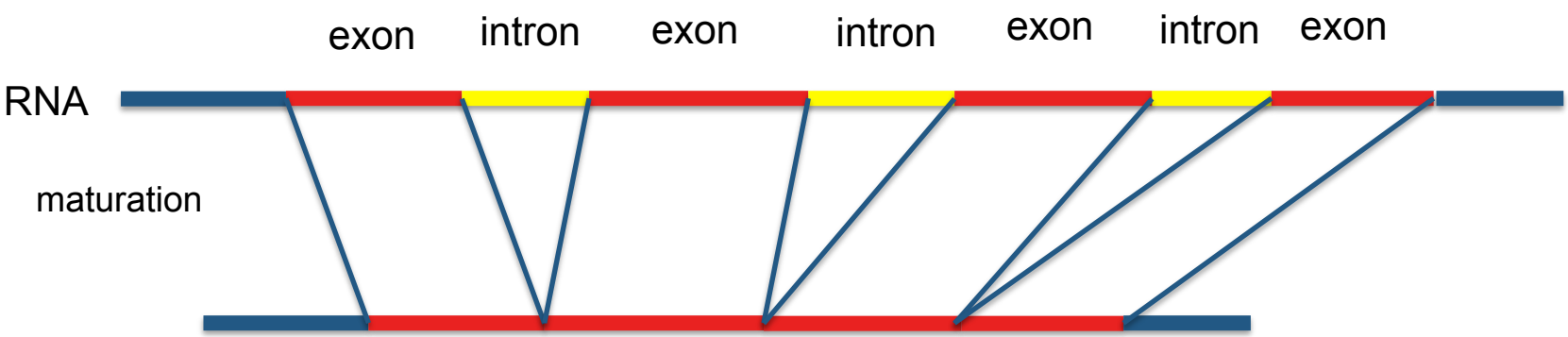
- Very long intergenic regions
  - Genes: less than 5% of a human genome

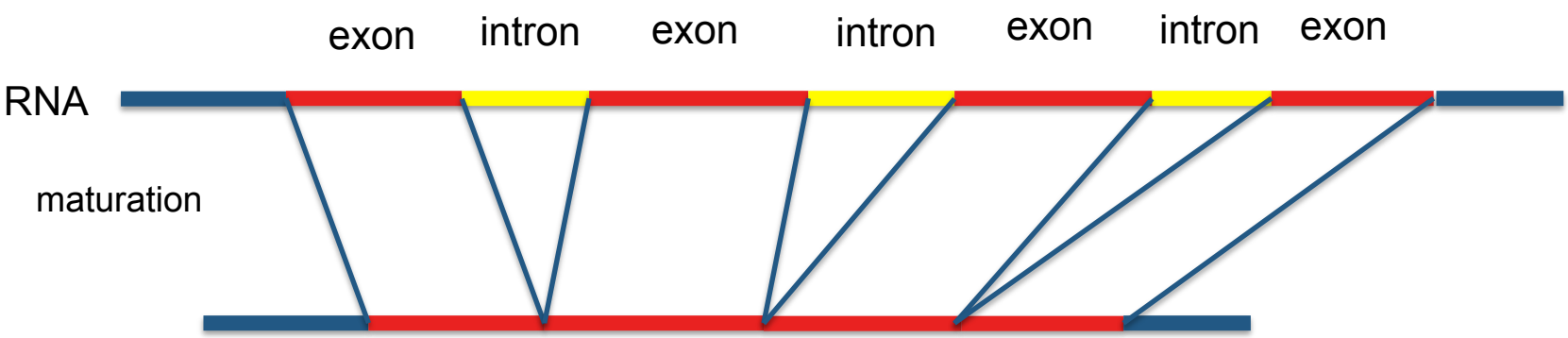
- Genes are interrupted by non-coding introns


exon
intron
exon
intron
exon
intron
exon
RNA

Eukaryotic genomes

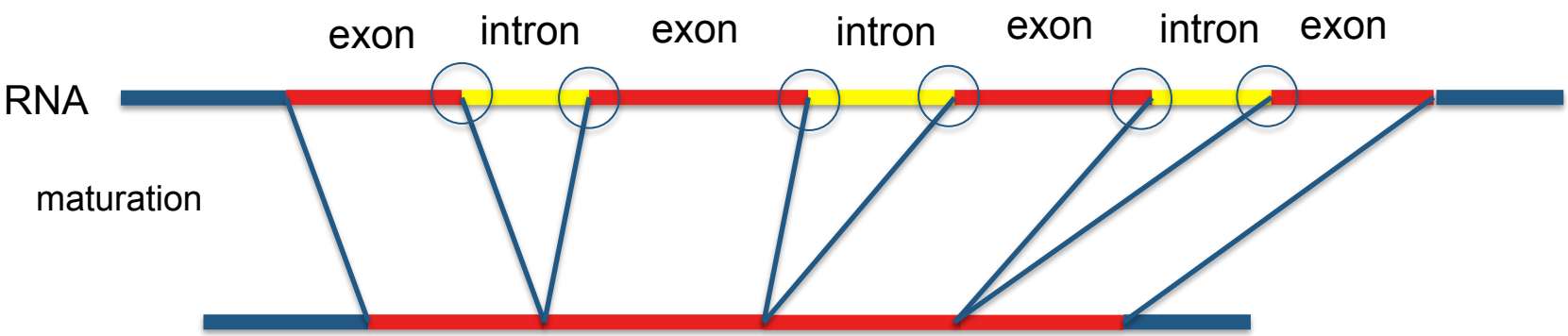
- Very long intergenic regions
  - Genes: less than 5% of a human genome
- Genes are interrupted by non-coding introns
  - Less then 2.5% of a human genome is made up of coding regions
- Introns are removed (spliced) during mRNA maturation


exon
intron
exon
intron
exon
intron
exon
RNA
maturation


exon
intron
exon
intron
exon
intron
exon
RNA
maturation

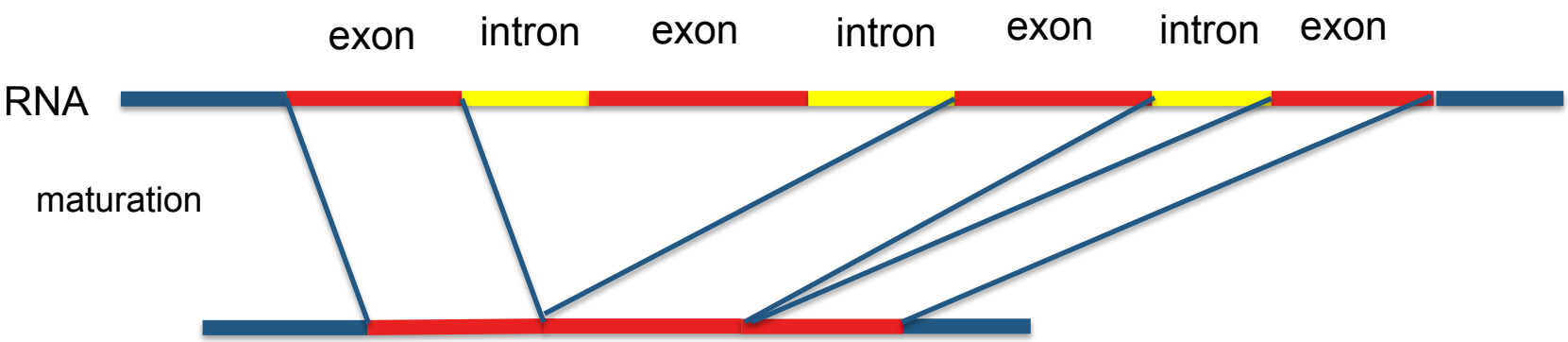
Eukaryotic genomes

- Very long intergenic regions
  - Genes: less than 5% of a human genome
- Genes are interrupted by non-coding introns
  - Less then 2.5% of a human genome is made up of coding regions
- Introns are removed (spliced) during mRNA maturation
  - Exon/intron junctions must be predicted


exon
intron
exon
intron
exon
intron
exon
RNA
maturation

Eukaryotic genomes

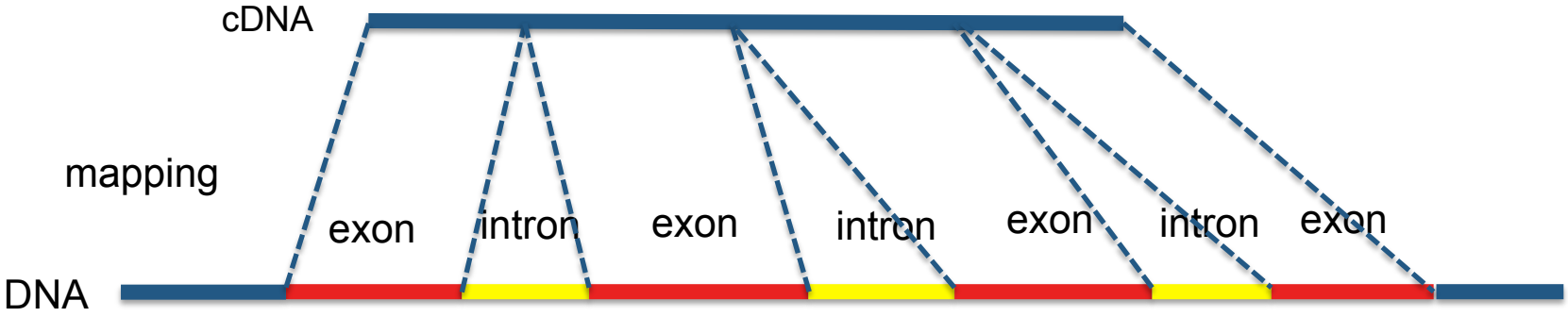
- Very long intergenic regions
  - Genes: less than 5% of a human genome
- Genes are interrupted by non-coding introns
  - Less then 2.5% of a human genome is made up of coding regions
- Introns are removed (spliced) during mRNA maturation
  - Exon/intron junctions must be predicted
- Alternative splicing
  - One gene, several proteins


exon
intron
exon
intron
exon
intron
exon
RNA
maturation

Use every source of knowledge

- Hidden Markov Models (HMM)
  - As many connected models as types of regions:

    coding, non coding, intergenic, etc.
- Patterns for exon/intron junctions
- Experimental data
  - complementary DNA (cDNA)


cDNA
mapping
exon
intron
exon
intron
exon
intron
exon
DNA

Still an open problem

- Eukaryotic gene prediction is still an open problem
- Gene predictors do not attempt to mimick the cellular processes
- Bioinformatic gene predictions are predictions