Annotation Workshop

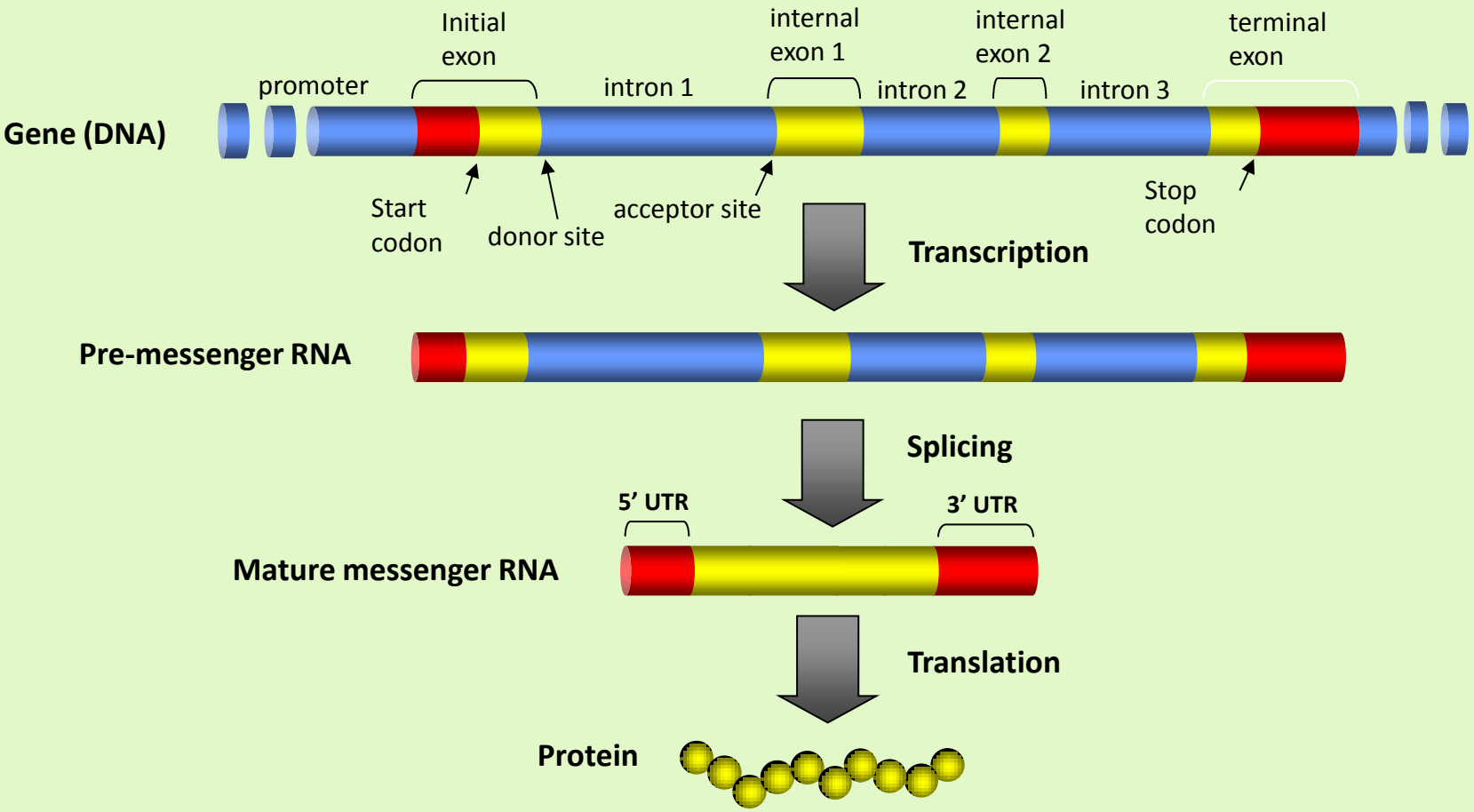
Lieven Sterck,
Bioinformatics & Systems Biology
VIB-UGent

Lieven.sterck@psb.vib-ugent.be

ProCoGen Dissemination Workshop, Riga, 5 nov 2013
"Conifer sequencing: basic concepts in conifer genomics"

"This Project is financially supported by the European Commission under the 7th Framework Programme"

Basic Eukaryotic Gene Structure

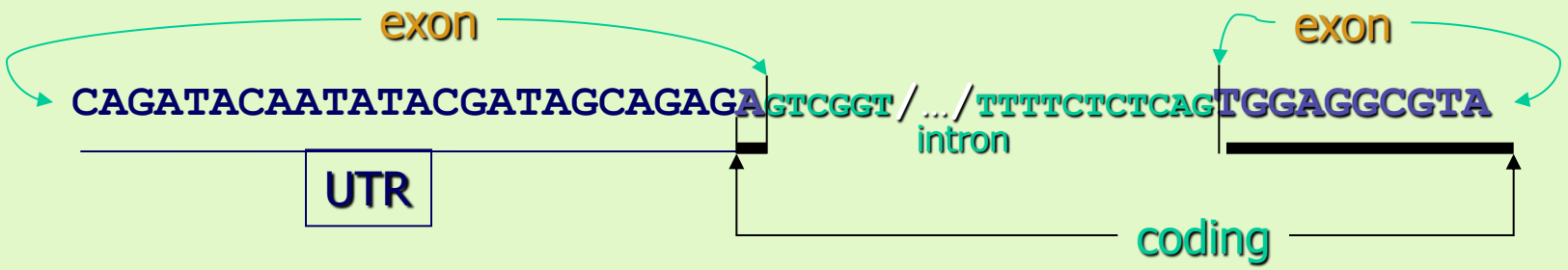


Basic rules

- Coding sequence (CDS) is modulo 3
- Starts with a start-codon ATG (M)
- Ends on a stop-codon (TAG, TAA, TGA)
- Usually consists of exons and introns
  - Intron starts with GT of GC dinucleotide
  - Intron ends with a AG dinucleotide
- No in-frame stop-codons

Points of attention

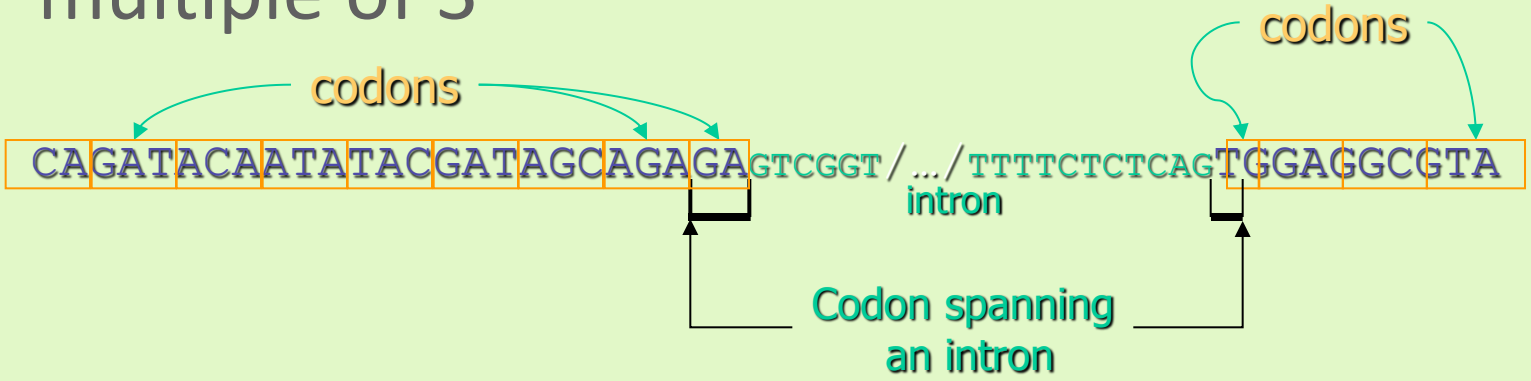
- Transcription, splicing and translation in eukaryotes are distinct processes



  - This means that CDS (coding sequences as reported in Genbank/EMBL files) can have an "exon" of size 1
    - True exon has UTR upstream of the CDS!
  - *ab initio g*ene-callers will be unable to predict those correctly need alignment to delineate exons correctly

Points of attention (2)

- Exons do not need to have a length equal to a multiple of 3



  - It is only after splicing is finished that the resulting open reading frame on the transcript has to be a multiple of 3
  - Take care for 'in frame stop-codons'

Links used in the demo

- Genomeview tool:

  http://genomeview.org/

- EuGene Annotation software:

  http://bioinformatics.psb.ugent.be/webtools/EuGene/