Assignment 1: Naïve Bayes Classifier of protein domains

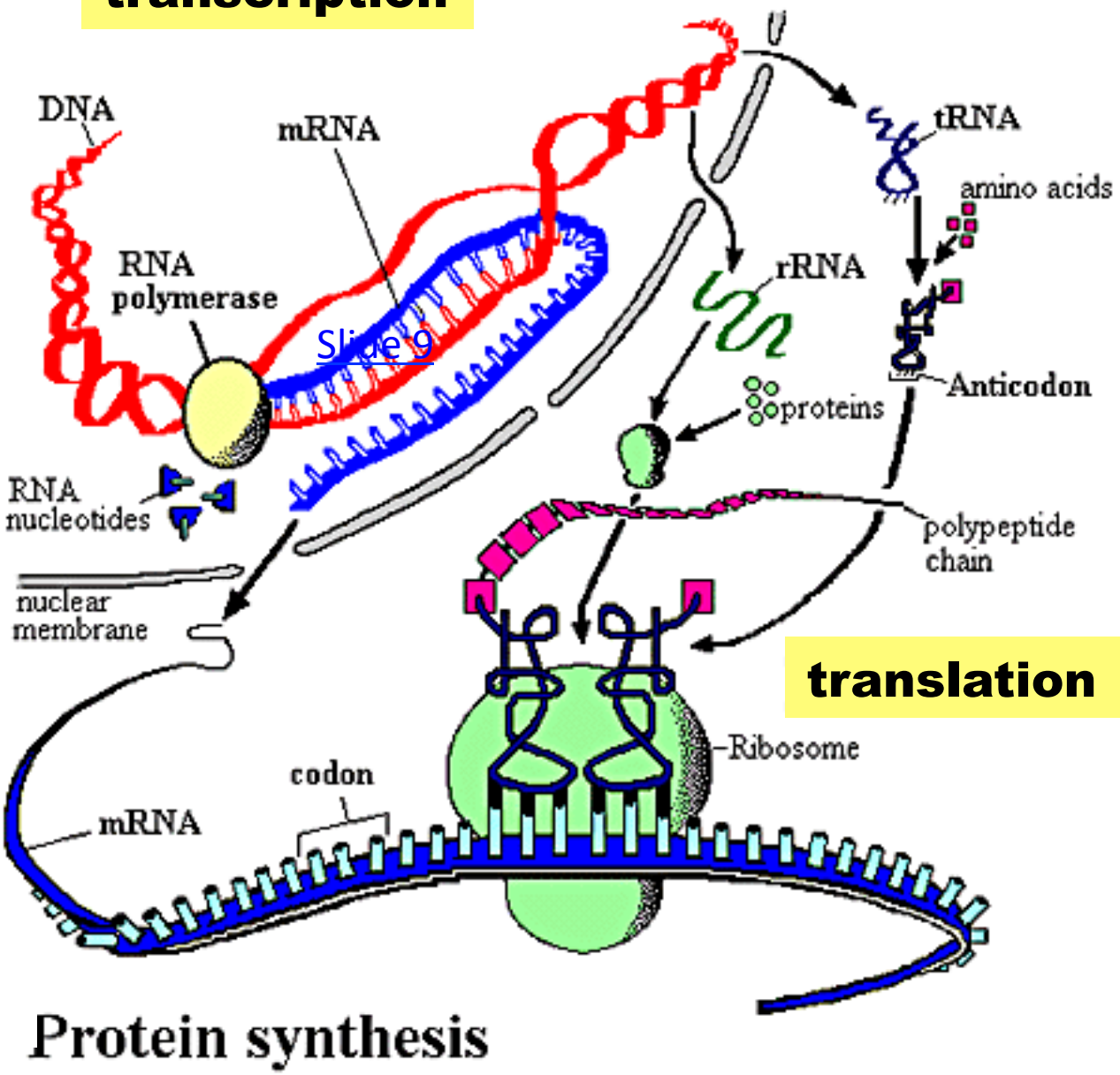
Master in Bioinformatics UPF
2010-2011

Eduardo Eyras
Computational Genomics
Pompeu Fabra University - ICREA
Barcelona, Spain

transcription

nucleus

cytoplasm

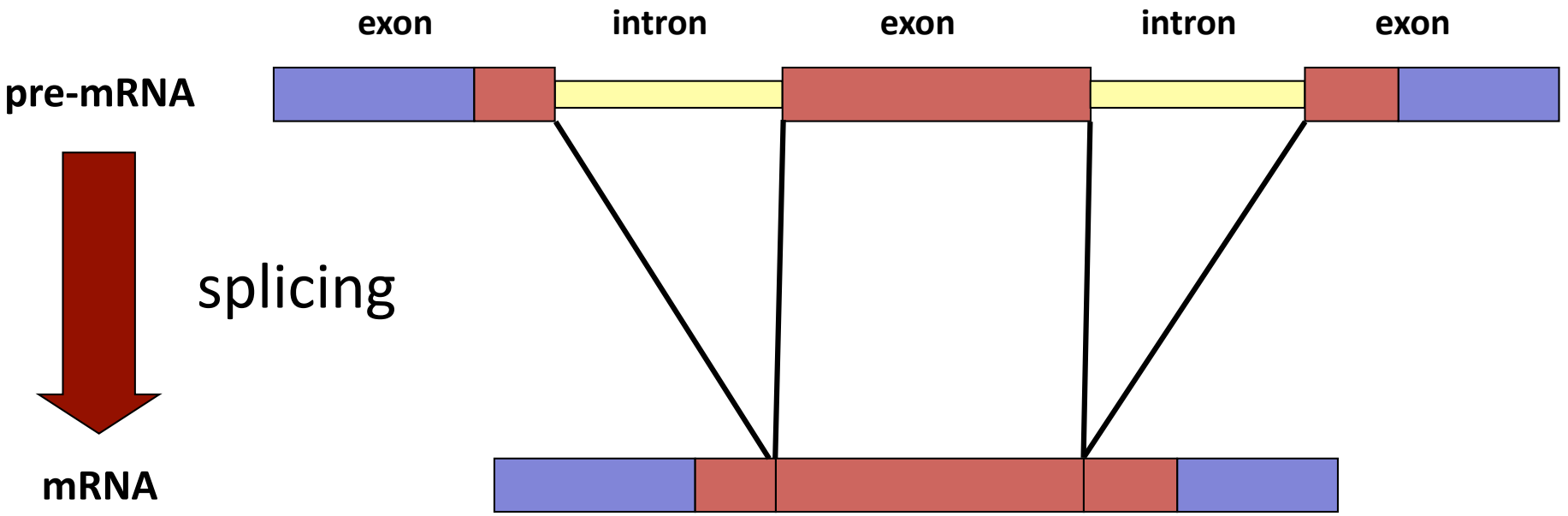


Slide 9

translation

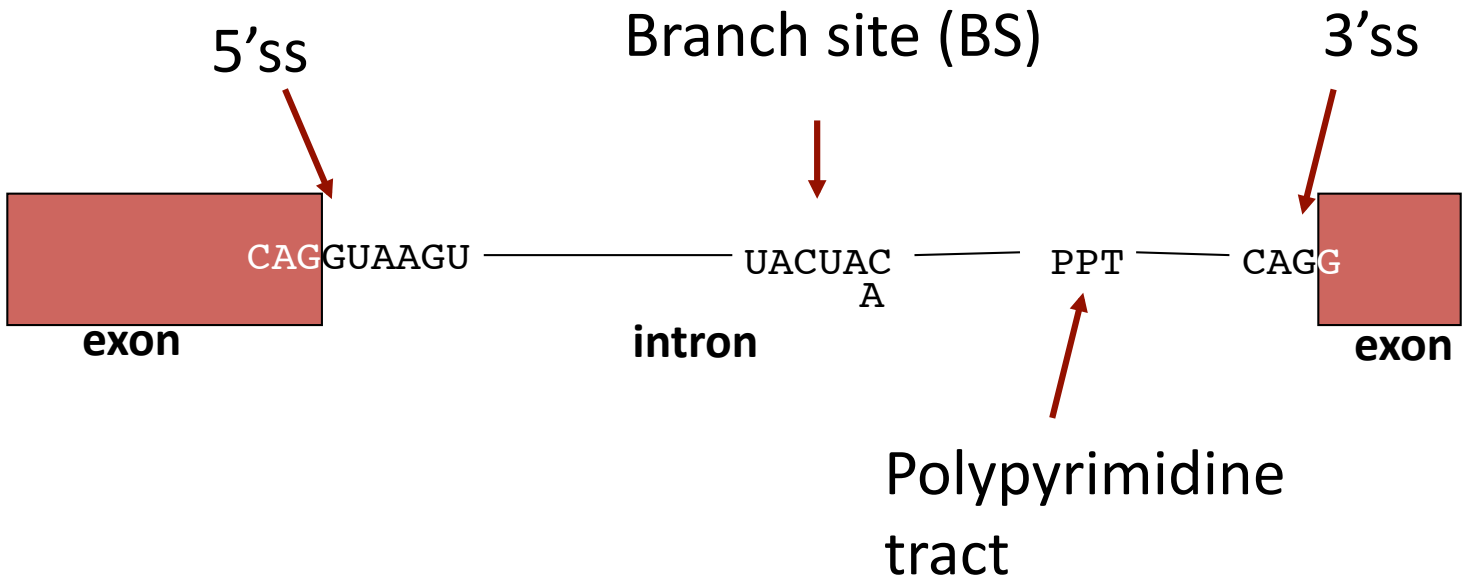
Protein synthesis

Splicing



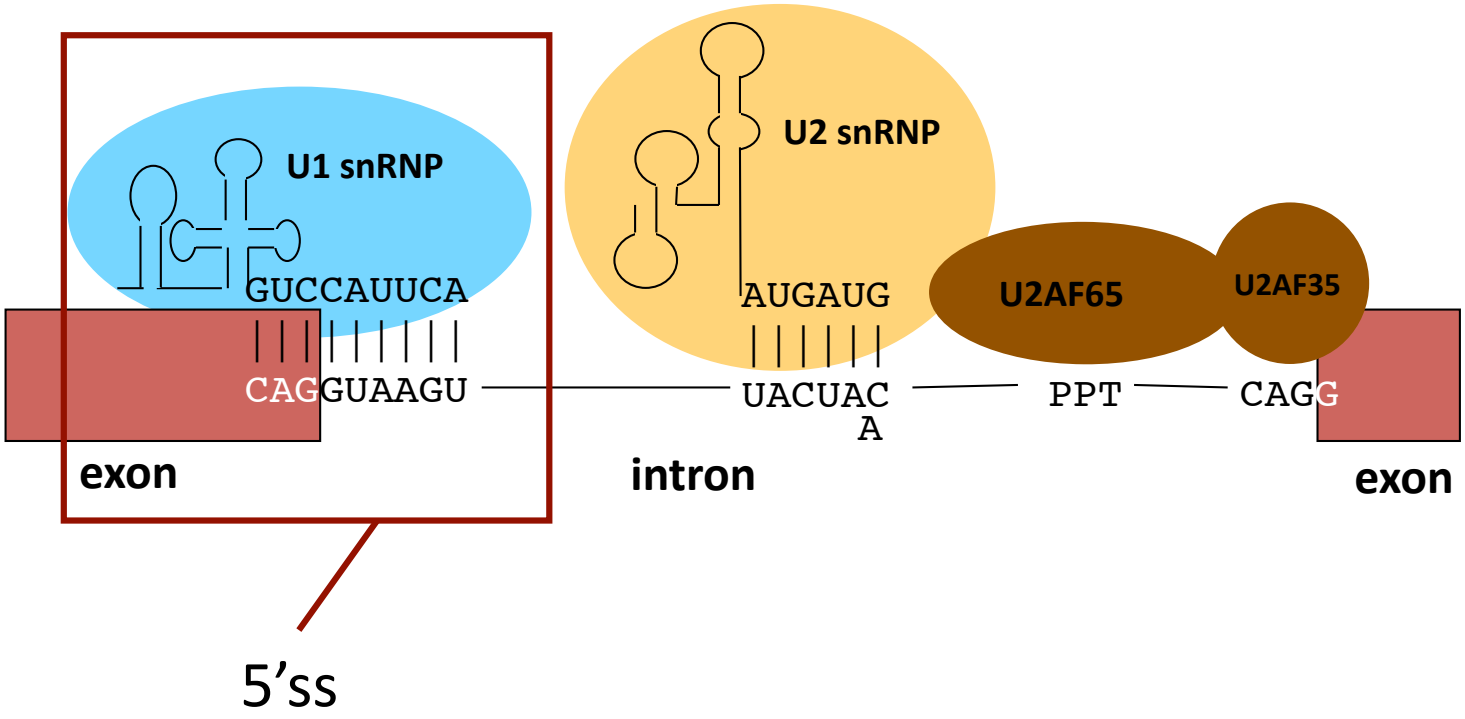
Introduction: splicing signals

4 basic splicing signals



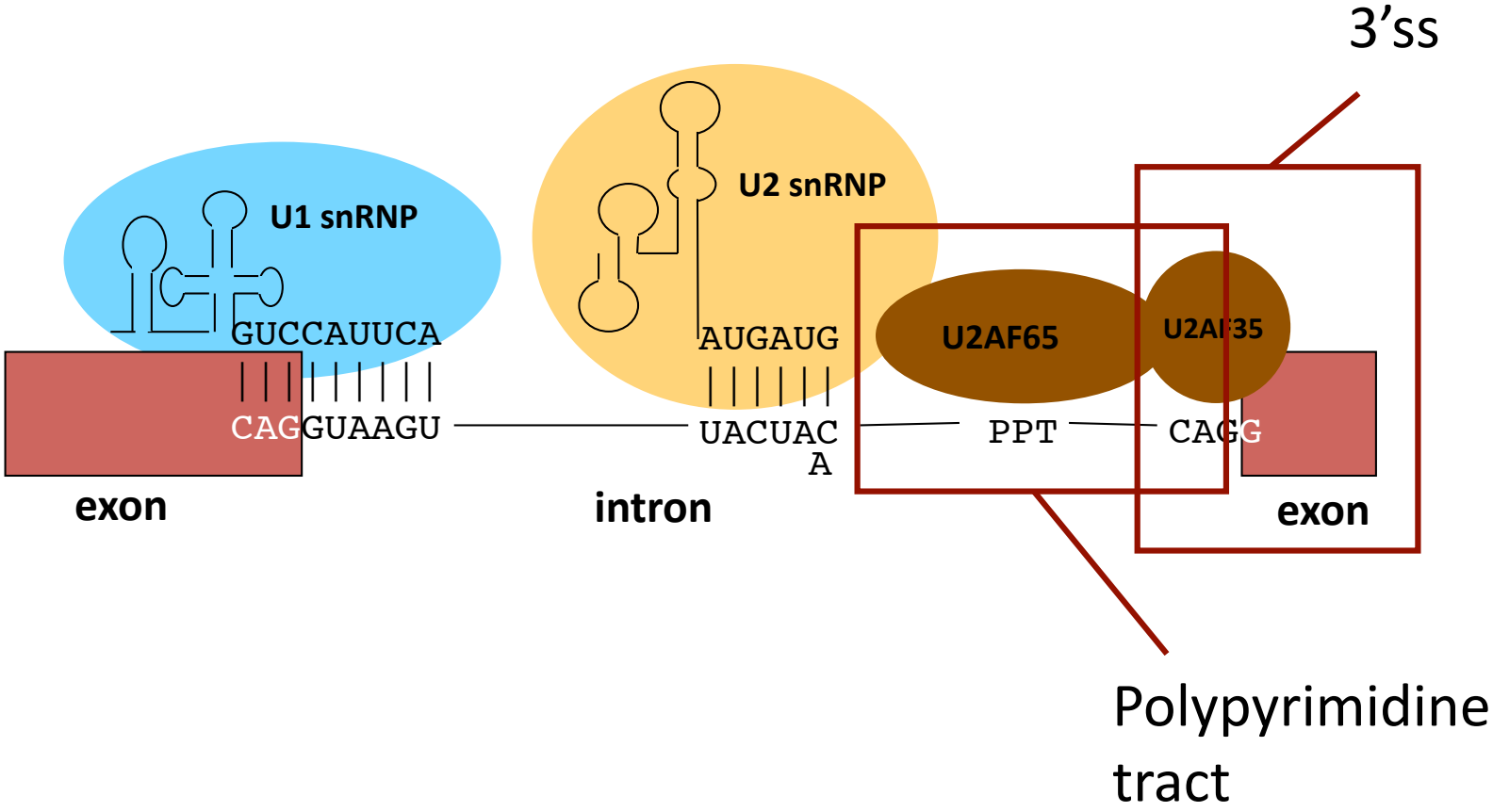
Introduction: splicing signals

4 basic splicing signals



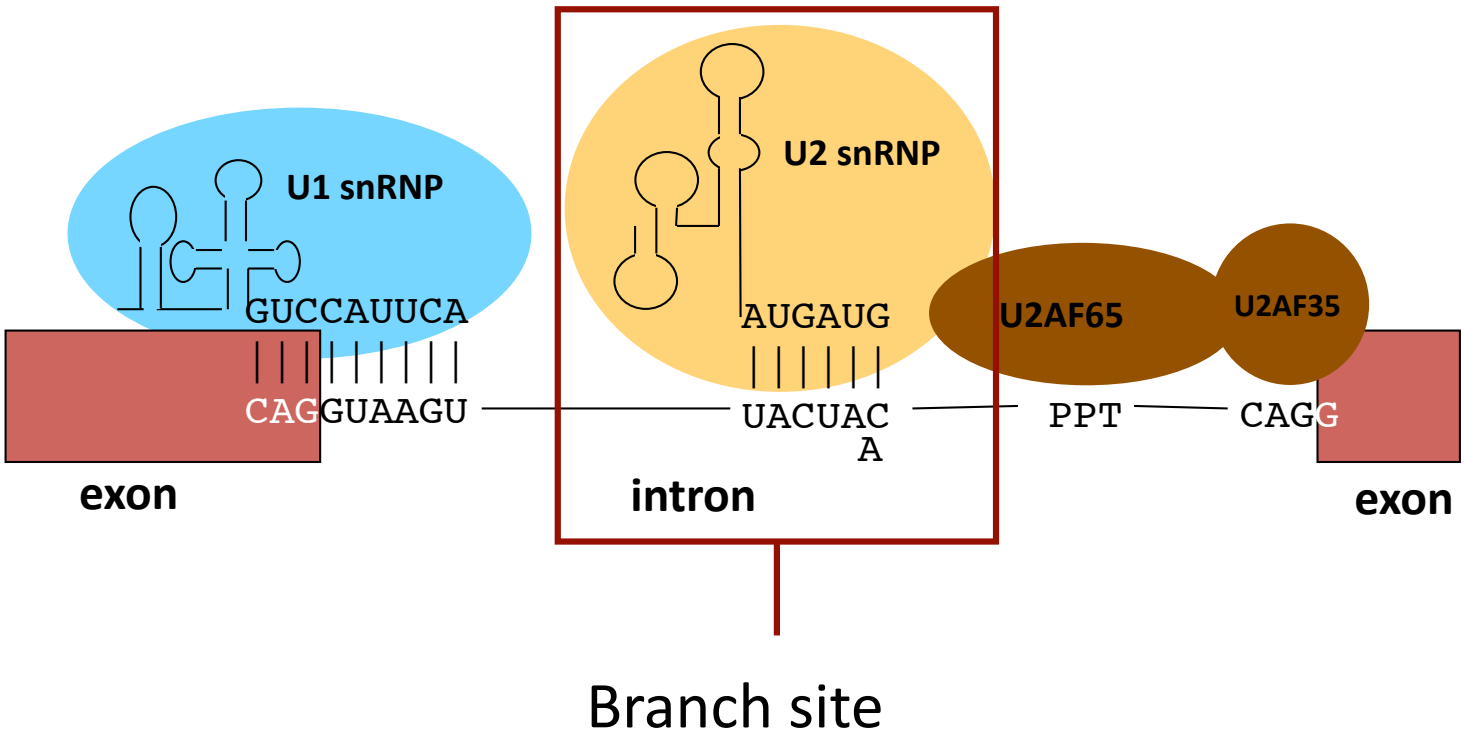
Introduction: splicing signals

4 basic splicing signals



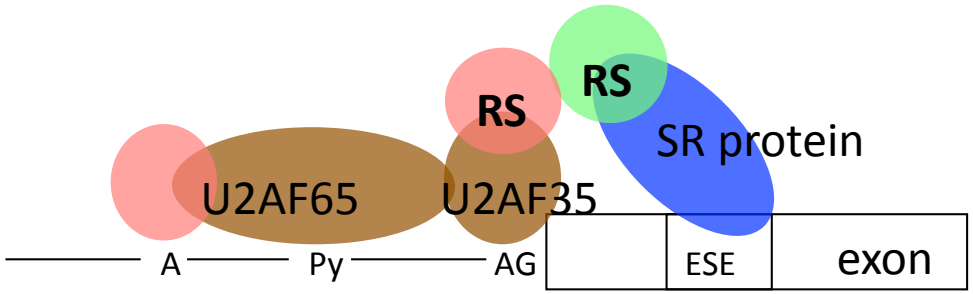
Introduction: splicing signals

4 basic splicing signals



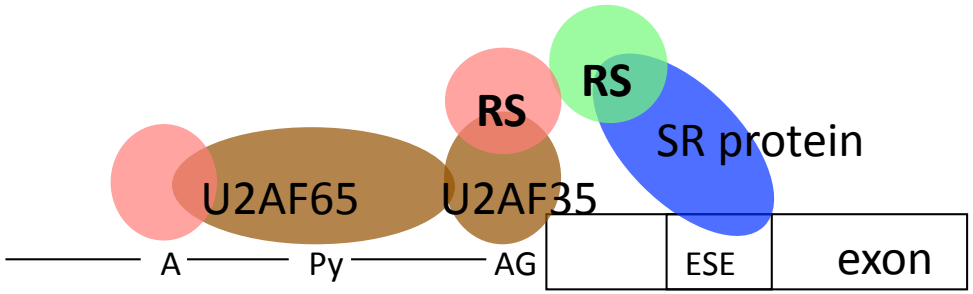
Splicing Factors (Protein-RNA interaction)

Splicing factors can bind to the pre-mRNA and promote (enhance) splice-site selection through the recruitment of spliceosomal components

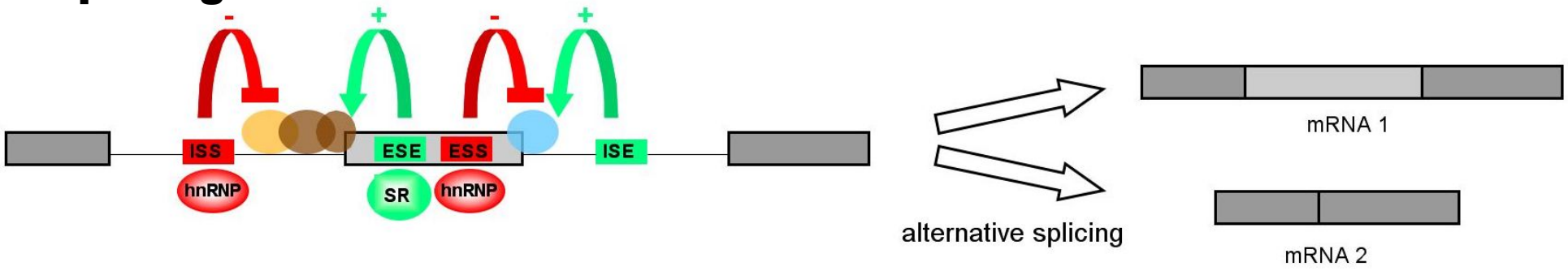


Splicing Factors (Protein-RNA interaction)

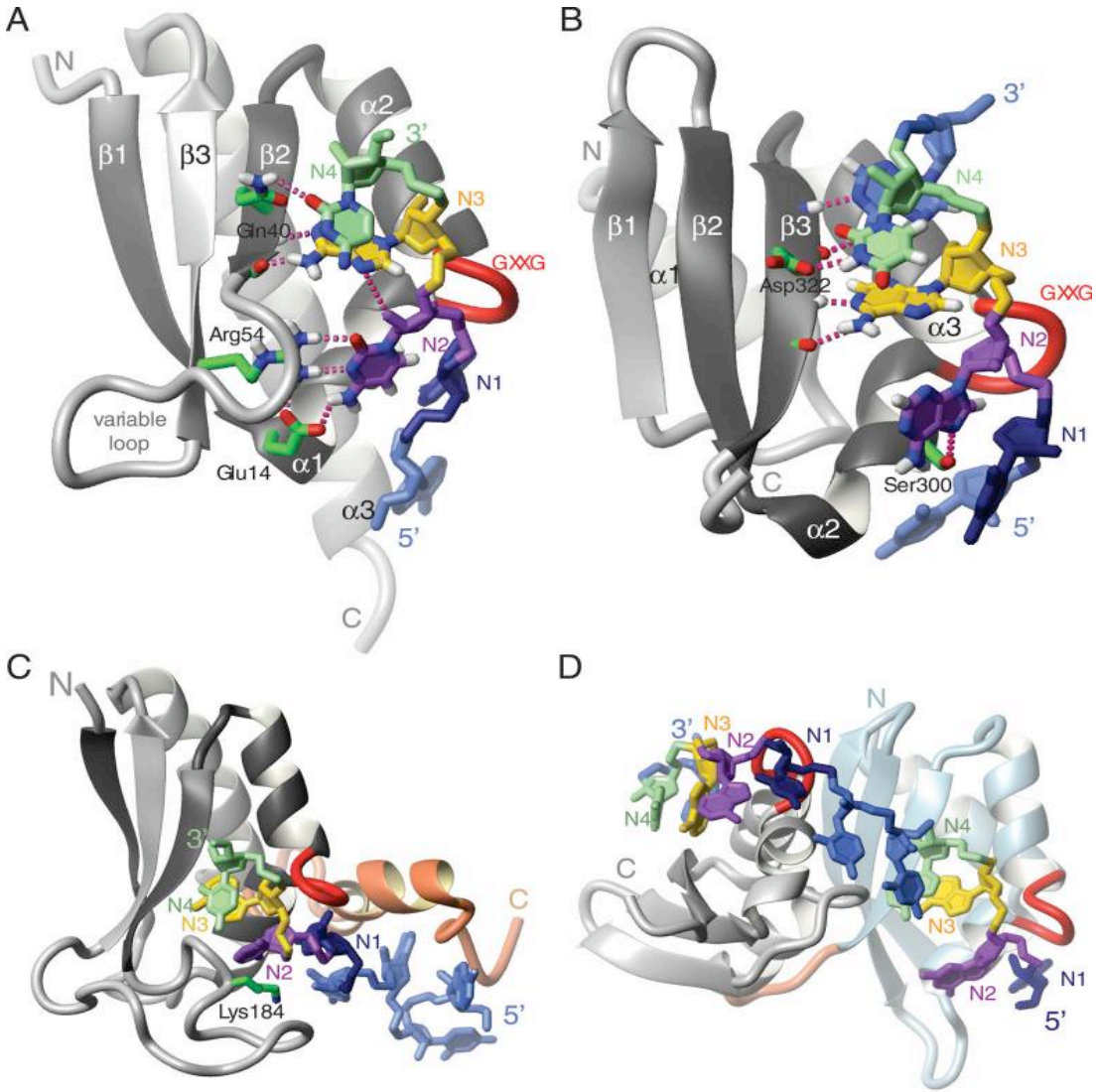
Splicing factors can bind to the pre-mRNA and promote (enhance) splice-site selection through the recruitment of spliceosomal components



In fact, the regulated action of factors gives rise to alternative splicing

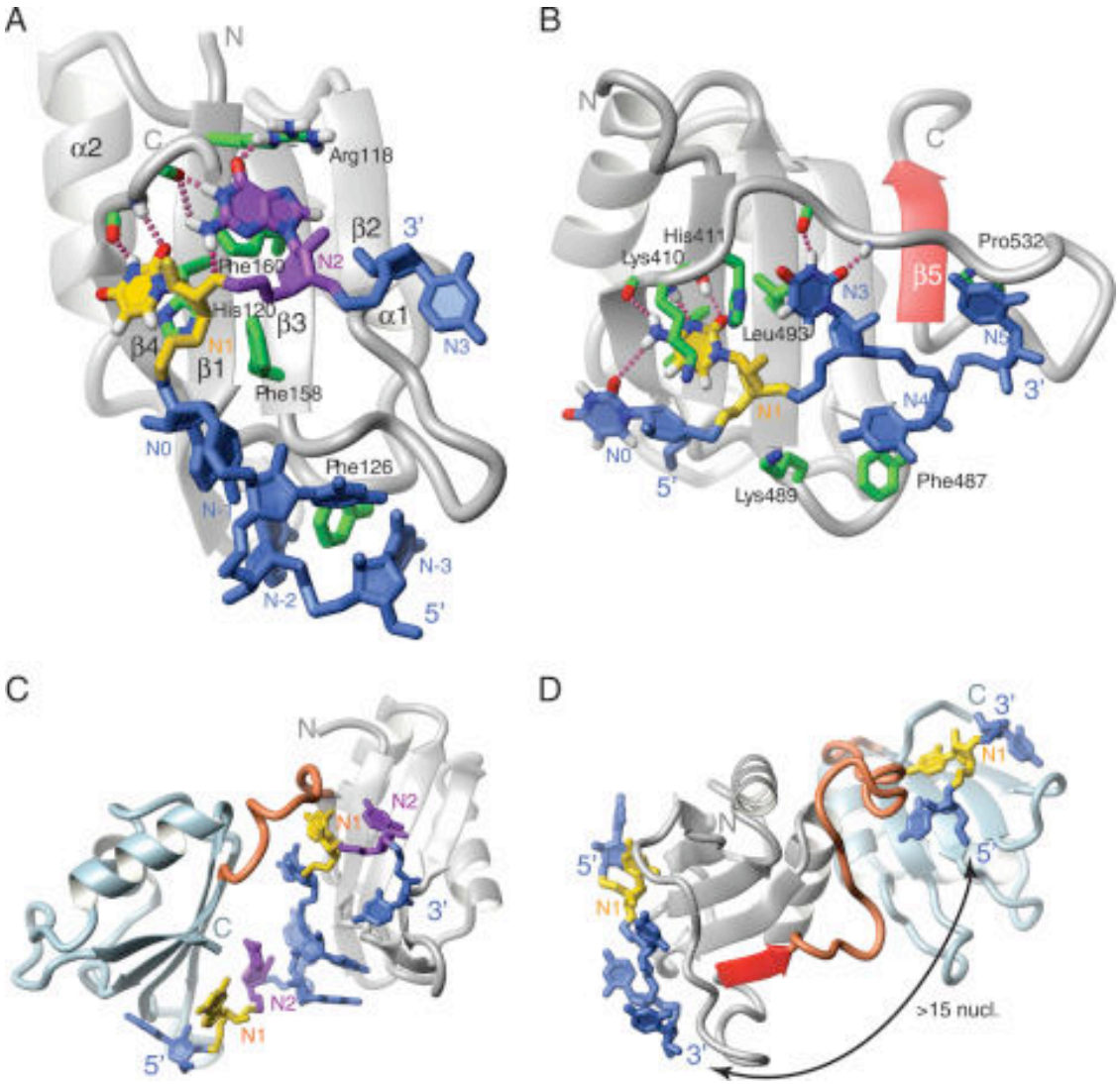


RNA Binding Domains



KH domains. (A) Type I KH domain of Nova (PDB code: 1EC6). (B) Type II KH domain of NusA (PDB code: 2ATW). (C) KH and QUA2 domains of SF1 (PDB code: 1K1G). (D) Tandem KH domains of NusA (2ATW)

RNA Binding Domains

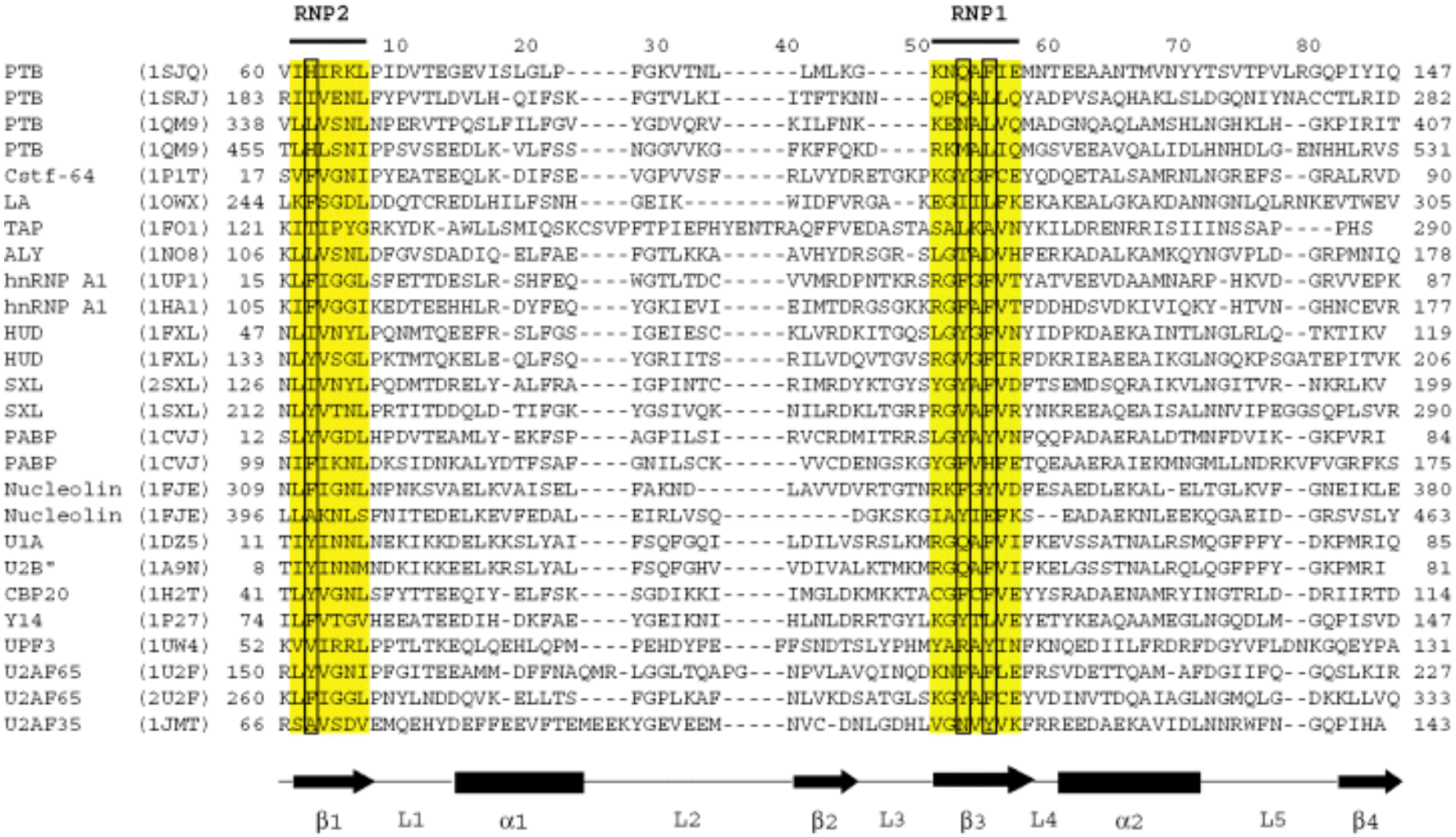


RRM domains. (A) The RRM of Fox-1 (PDB code: 2ERR). (B) RRM3 of PTB (PDB code: 2ADC). (C) The tandem RRMs of Sex-lethal (PDB code: 1B7F). (D) RRMs 3 and 4 of PTB (PDB code: 2ADC)

The RNA Recognition Motif (RRM)

Hexamer RNP2 = [IVL] [FY] [IVL] X N L

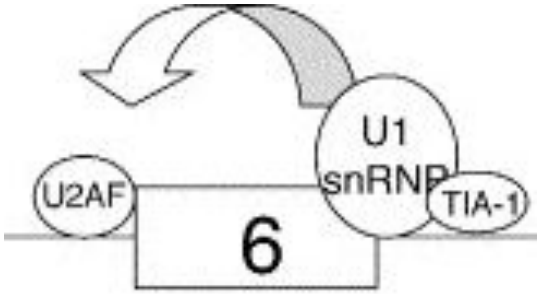
Octamer RNP1 = [LR] G [FY] [GA] [FY] [VIL] X [FY]



Aromatic residues for primary RNA binding: F, Y, (W,H)

Other signatures: DAEDA (not found in all SR/SR-like proteins)

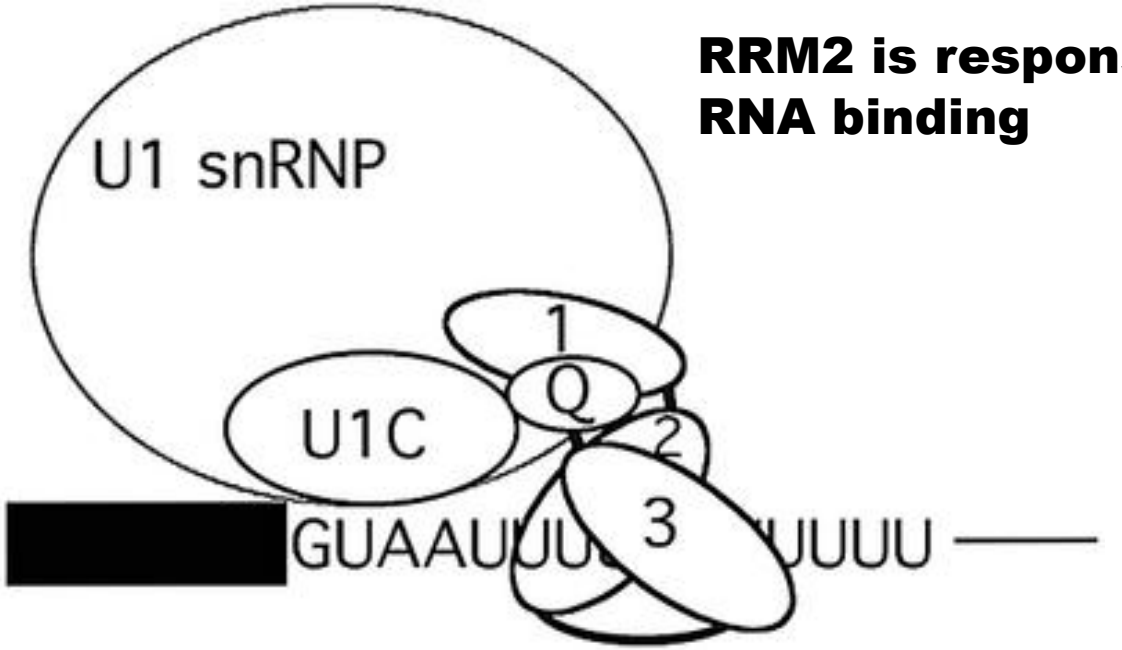
TIA1 can recruit U1 snRNP



Izquierdo et al. 2005

TIA1

RRM2 is responsible for RNA binding



Forch et al. 2002

TIA1 is conserved across eukaryotes

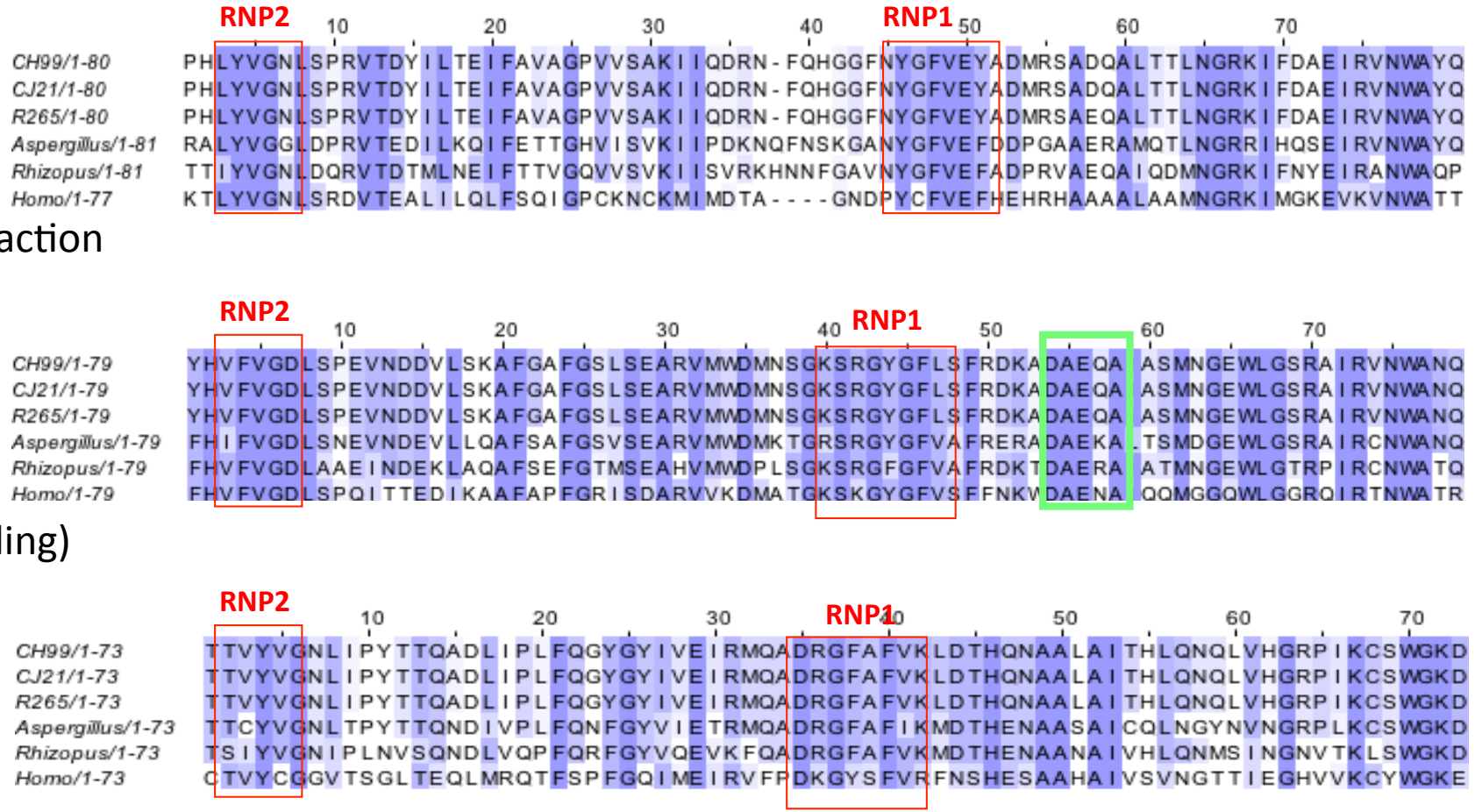
RRM1

(U1) interaction

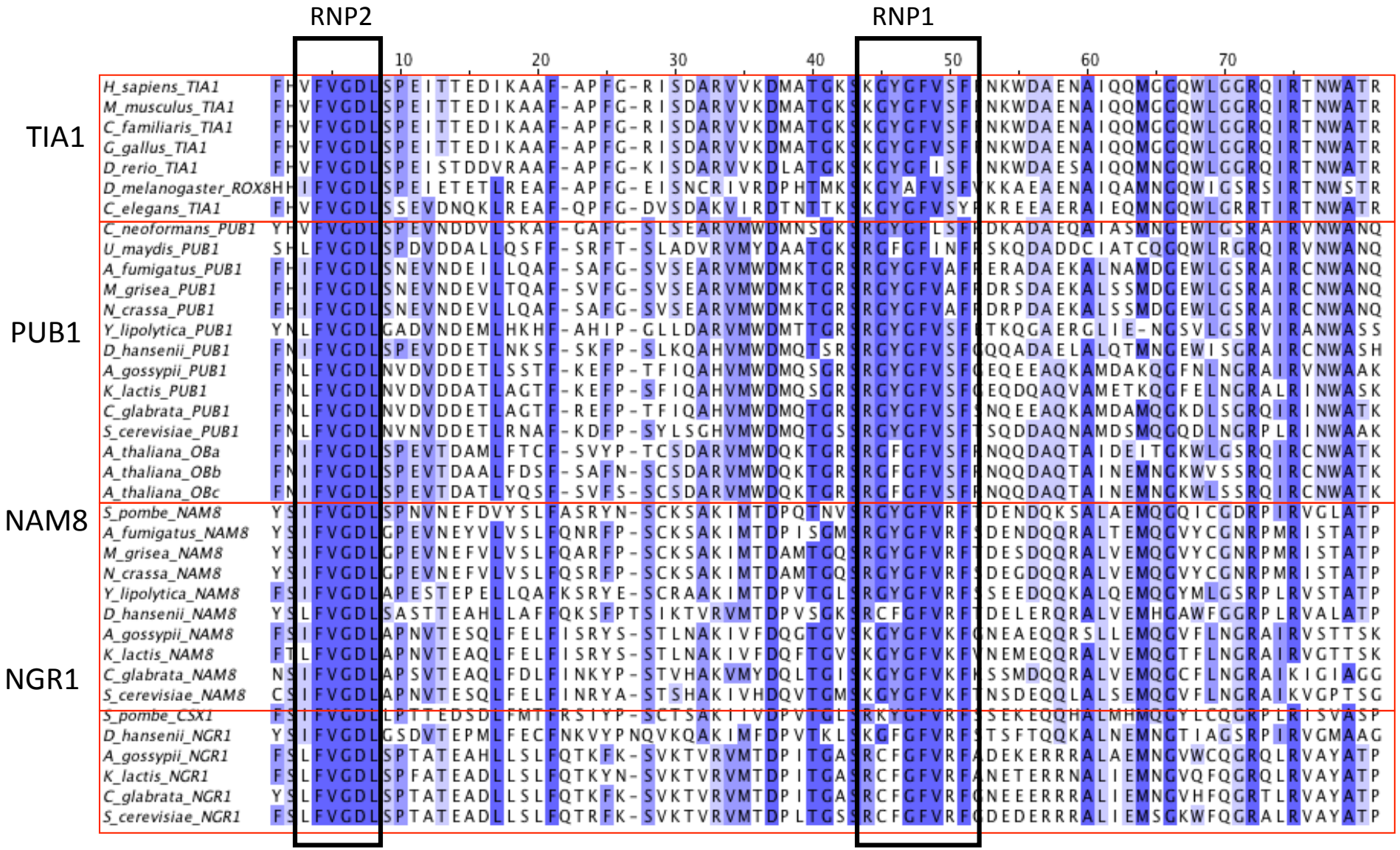
RRM2

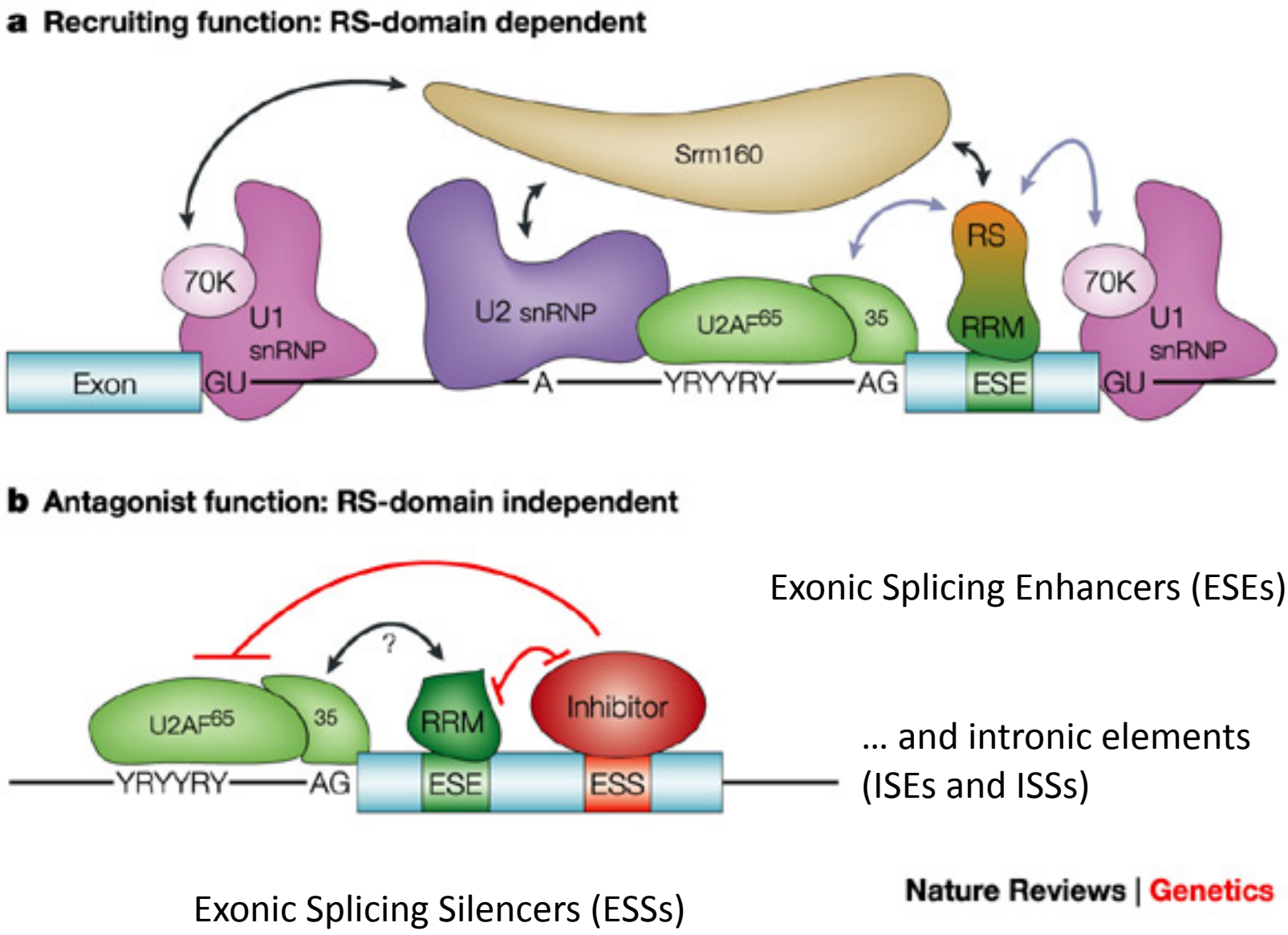
(RNA binding)

RRM3



Functionally related proteins have also conserved RRM domains




a Recruiting function: RS-domain dependent
Srm160
70K
U1
snRNP
U2 snRNP
U2AF65
35
RS
RRM
70K
U1
snRNP
Exon
GU
A
YRYYRY
AG
ESE
GU
b Antagonist function: RS-domain independent
?
U2AF65
35
RRM
Inhibitor
YRYYRY
AG
ESE
ESS
Exonic Splicing Enhancers (ESEs)
… and intronic elements
(ISEs and ISSs)
Exonic Splicing Silencers (ESSs)
Nature Reviews | Genetics

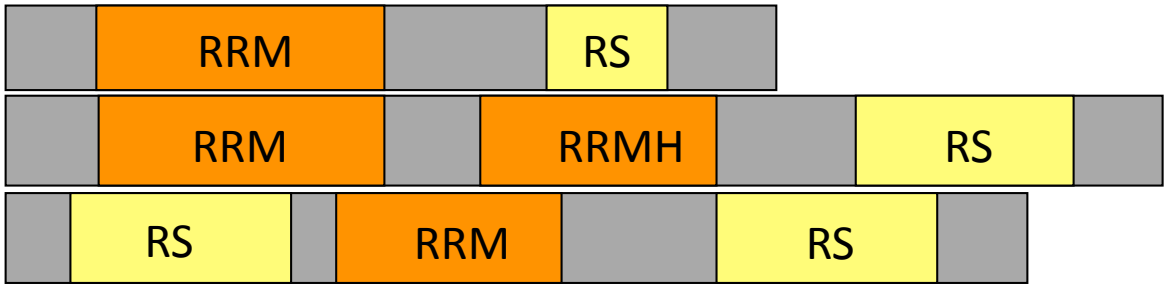
Domain organization of splicing factors

SR proteins

SC35, SR46, SRp20, 9G8, p54, SR86

SRp40, SRp55, SRp75, ASF, SR30C

TRA2, RNPS1



hnRNPs

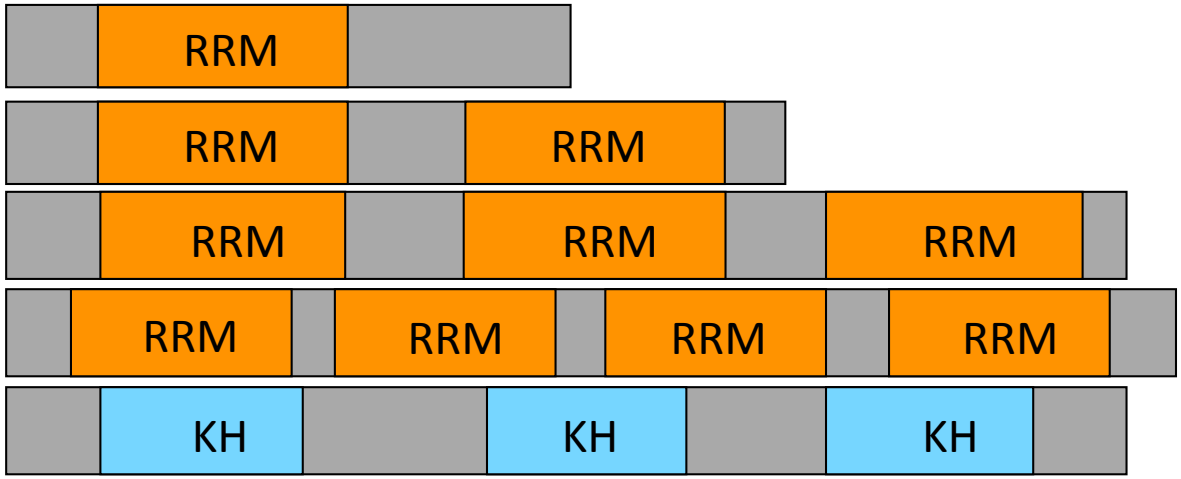
hnRNP-C, hnRNP-D, hnRNP-G

hnRNP-A, hnRNP-D, hnRNP-F/H, musashi

hnRNP-L, hnRNP-M, hnRNP-R

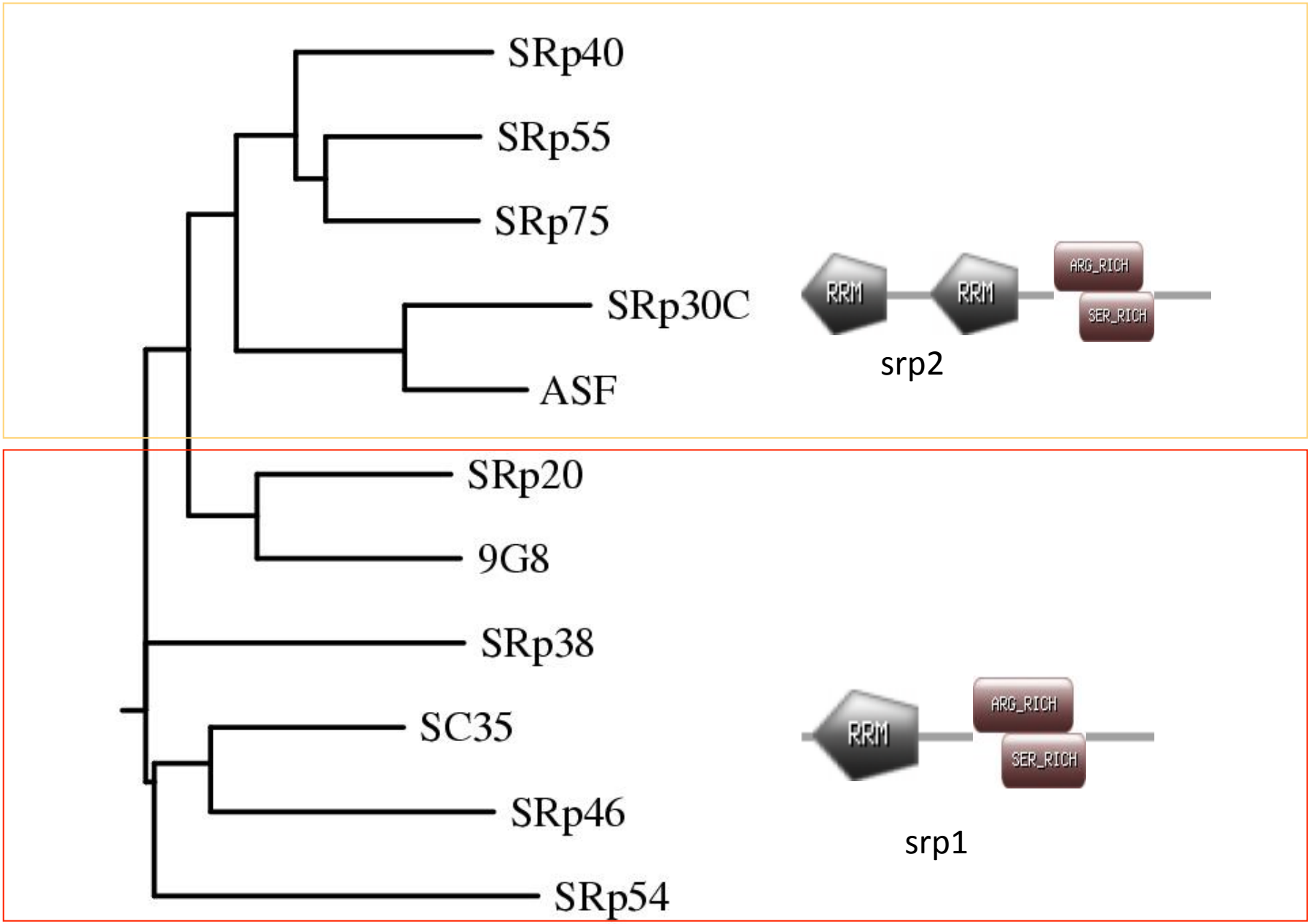
hnRNP-I

hnRNP-E



RRM: RNA Recognition Motif
RRMH: RRM homologous (similar to RRM)
RS: variable Arginine/Serine rich region (contains RS or SR dipetides)
KH: K-Homology domain

Human SR proteins



Starting hypotheses

Proteins with the same evolutionary origin (orthologs) are conserved
(have conserved functional sites)

Proteins with similar functions and related evolution (homologs) are also conserved

Proteins that perform similar interactions (Protein-RNA) but have opposite function
Are they conserved??

Can we model these differences/similarities with a simple model?

Objective

Build a Naïve-Bayes model with information from three types of data:

RRMs from enhancer proteins: SR proteins
RRMs from repressor proteins: hnRNP
KH-domains from repressor proteins: hnRNP
RRMs from other splicing factors (non-SR, non-hnRNP)

Can this model correctly classify a list of unknown domains?