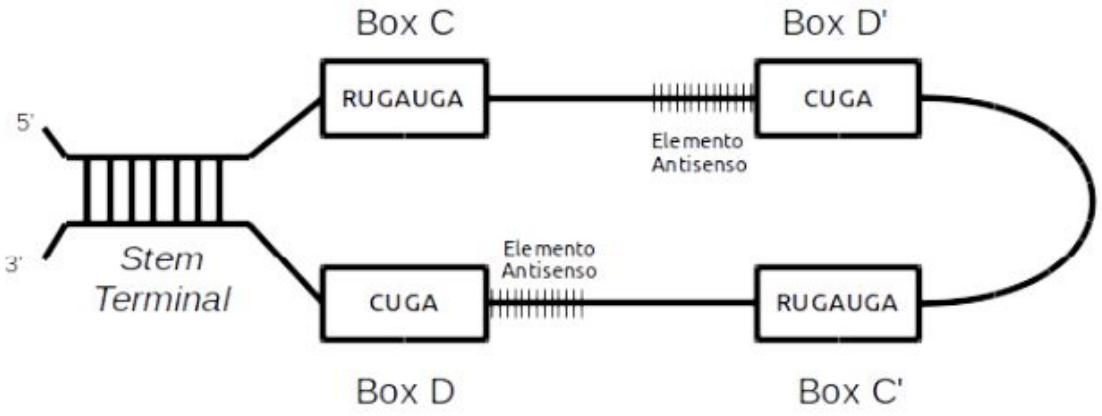
Instituto Federal de Brasília - Campus Taguatinga
Superior em Computação

PB 4
Busca de Motivos em Sequências Conservadas

Introdução à Bioinformática 1/2019
Professor João Victor de Araujo Oliveira

Motivos conservados em C/D box snoRNAs

C/D box snoRNAs são caracterizados pela presença de dois motivos conservados, o box C (RUGAUGA) e o box D (CUGA).



Os arquivos de texto C.motif e D.motif mostram uma coleção de sequências de boxes C e D conhecidas, respectivamente. Nestes arquivos, cada linha expressa uma sequência única de box conhecida e a quantidade de ocorrência dela em diferentes organismos vertebrados (veja a figura abaixo).

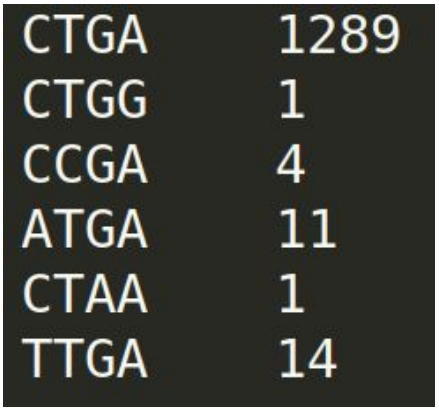


Figura: Arquivo D.Motif: contém em cada linha uma sequência única do box D e a quantidade de ocorrências que ela ocorre nas amostras disponíveis.

Exercício: Crie um programa que leia os arquivos C.motif e D.motif e crie uma PSSM para os boxes C e D de um snoRNA. O programa deve implementar as seguintes funcionalidades (para cada box):

1. Exibir no terminal uma matriz com as frequências de cada resíduo em cada posição do alinhamento múltiplo similar a apresentada em sala de aula (figura abaixo);

Pos.	1	2	3	4	5	6	Overall freq.
A	0.6	0.6	—	0.4	—	0.2	0.30
T	0.2	0.2	—	0.4	0.2	0.2	0.20
G	—	0.2	0.6	—	0.2	0.6	0.27
C	0.2	—	0.4	0.2	0.6	—	0.23

2. Exibir no terminal uma matriz com as frequências normalizadas, dividindo as frequências posicionais de cada resíduo pela frequência total;

3. Exibir no terminal uma matriz com os scores normalizados para escala logarítmica na base 2;

4. Por fim, deve ser salvo em um arquivo tabular com os scores totais de cada amostra contida nos arquivos C.motif e D.motif.

Obs.: Pode ser um único programa que exibe as matrizes e salva os scores nos arquivos C.motif.scores e D.motif.scores.