Finding splicing signals in the PC gene

Strategy:

1. Open UCSC genome browser and search for human PC gene
2. Click on „Tools -> Table Browser“
3. Tick „postion“ in the region section
4. Select „sequence“ as „output format“
5. Click „get output“ and select „genomic“, click submit
6. Select „CDS Exons, 5‘UTR Exons, 3‘UTR Exons“ and „One FASTA record per region (exon, intron, etc.) with extra 10 bases upstream (5') and extra 0 downstream (3')” AND vice versa
7. Check “Exons in upper case, everything else in lower case”
8. Copy the output to Notepad++

Regex keeping the 3‘ or 5‘ splice sites and remove anything else

[…]
>hg19_refGene_NM_000920_0 range=chr11:66725783-66725847
GTCAGTGGAGGCAGCAGCGGTAGAGGCGGCGGCGAGGACTGGCGACGGCGAGGAGgtgagtgtcc
[…]

Or

[…]
>hg19_refGene_NM_000920_0 range=chr11:66725793-66725857
gctgttctctGTCAGTGGAGGCAGCAGCGGTAGAGGCGGCGGCGAGGACTGGCGACGGCGAGGAG
[…]

TO

CGGCGAGGAGgtgagtgtcc

OR

gctgttctctGTCAGTGGAG

Regex keeping the 3‘ or 5‘ splice sites and remove anything else

1. Get rid of the “fasta headers” using search replace with the following regex

   Search “**^>.+**” and replace with “”

2. Get rid of the line breaks at the sequence blocks:

   search “**(.+)($\r\n)**” replace “**\1**”
   (use only “\n” if formatting is UNIX)

3. Keep only the first 20 or last 20 bps as single lines:

   search “**(^.{20})(.+)**” replace “**\1**”

   Or

   search “**(.+)(.{20}$)**” replace “**\2**”

Use these sequences to generate a Sequence Logos at http://icbi.at/logo/