**RAAPI: Reduced amino acid pattern identification**

**Executable code file:**

There is a single executable code file (***RAAPI***) written in python script. One needs to download the executable code file, which can be executed in any python (version >=3.0) installed computer. Once executed (by click only), program will ask for several inputs parameters and then it automatically generate output files as described below.

1. **Input parameters: (once execute the code, program will ask for five input parameters below)**
2. Targeted group protein sequences in fasta format (say, *Tg.fasta*). This is the main fasta file where RAAPI pattern will be searched by the program.
3. Remaining group protein sequences in fasta format (say, *Rg.fasta*)
4. Searching upper pattern length (pu), say, *25*
5. Searching lower pattern length (pl), say, *10*
6. Effectiveness score threshold (Et), say *20*

\*\*Input should be file name only without *.fasta* (first two points) or value only (last three points) (as highlighted above)

\*\*Must keep both the fasta format input files (***Tg and Rg***) withing the same directory where main program file is being executed.

\*\*Once execute (by click) ***RAAPI*** file, a message window will generate to provide all five input values, once provided, just click on ***OK***. The console output can be seen for different steps status.

Effectiveness threshold (Et) score is the percentage of protein sequences having same pattern (already found in ***Tg***) from ***Rg*** group protein sequences.

**Important points to be noted:**

1. If interest is only for searching common pattern in targeted group sequence (***Tg***), provide both files name at points 1 and 2 above in A as ***Tg*** (same). (In this case no Ef score in output files see below B)
2. Must be ***pu>pl*** (points 3 and 4 above in A)
3. Automatically pattern search start from higher length (***pu***) to lower length (***pl***). So, lower length pattern is discarded if it is a subsequence of already found common higher length pattern.
4. If needs to search only specific pattern length, then provide pattern length ***pu=pl*** (points 3 and 4 above in A)
5. **Output files:**
6. Only common patterns list with pattern length (saved as ***Tg\_result1***)
7. Numerical pattern with Ef score, sequences ids (saved as ***Tg\_result2***)
8. Numerical pattern with Ef score, position, and amino acid sequence for each sequence (saved as ***Tg\_result3***)

\*\*For the last two files, pattern effectiveness (Ef) score is the percentage of protein sequences having same pattern (already found in ***Tg***) from ***Rg*** group protein sequences. And percentage cutoff is based on effectiveness threshold provide at point 5 in A.

**General comments:**

Program usually takes less than 45s to generate and identify and saving common pattern list for around 100 sequences (25 targeted group protein sequences and 75 remaining group protein sequences) of each length about 200-300 amino acids.