Basic functionality of software\*:

1. User input nucleotide sequence corresponding to an ORF, or amino acid sequence
2. Options to select the categories of hotspots that are to be optimized (see below)
3. Algorithm that:
   1. Translates the ORF into an amino acid sequence
   2. Searches the nucleotide sequence for incidences of selected hotspots
   3. Determines where nucleotide substitution(s) can be done that would eliminate a given hotspot while not changing the amino acid sequence if the user selected for sense substitution, *or only make conservative changes in amino acid sequence if the user selected for conservative substitution.*
   4. *If the user selected to increase mutation potential, searches through nucleotide sequence to find where nucleotide substitution(s) can be done that would create a hotspot while not changing the amino acid sequence if the user selected for sense substitution, or only make conservative changes in amino acid sequence if the user selected for conservative substitution*
   5. *Provides the option for the user to mark parts of the sequence as unchangeable, for instance when conservative is selected with a protein containing a sensitive active site*
   6. *Option for user to input their own DNA sequence that the program will then look for and, according to weight user assigns, get rid of or enhance*
4. Outputs sequence, visibly highlighting what regions were changed from the original *and displaying what type of hotspot was eliminated, and if conservative, what amino acid has taken its place.*

\*Portions not in italics are essential for a minimum program. Italicized portions are added functionalities that can be done after the basic functionality is established.

Hotspot list (subject to change)

* Simple Sequence Repeats
  + Same nucleotide *or pair/triplet of nucleotides* repeated more than four times in a row, *or based on a different number of repeats the user can input*
* Homologous Repeats
  + Any nucleotide sequence of greater than five bases *or a different number the user can input* that appears more than once in the sequence, including as an inverted repeat
* E. coli Insertion Sequences (https://www-is.biotoul.fr/)
  + TGCGGACGATCATCAGTTAT (ISEc17)
  + GATCGTTGGGAACCG (IS903B)
  + GCAGTCAGGCACCGT (IS50R)
  + TAAGCTTTAATGCGC (IS50R)
  + GCAGTCAGGCACCGT (IS50R)
  + GCCGCCCAGTCCTGC (IS50R)
  + GTCTGACGC (IS50R)
  + CAGGAAAGA (IS3411)
  + AGTGCCATCTCCTT (IS30)
  + TTACCTGGTGC (IS30)
  + TTGTGTTTTTCAT (IS1A)
* E. coli Methylase Sites
  + GATC (Dam)
  + CCWGG (Dcm)
* Miscellaneous Hotspots (see Rogozin and Pavlov 2003 ; Cooper et al 2011)
  + CG (deamination)
  + RG (alkylation)
  + YY (ultraviolet)
  + YTC (ultraviolet)
  + GGA (oxidation)
  + YTG (polymerase)
  + GTGG (spontaneous)
  + SM (polymerase)
  + RGYW (spontaneous)
  + CNG (methylation)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Symbol** | **Description** | **Bases represented** | | | | |
| **A** | **a**denine | A |  |  |  | 1 |
| **C** | **c**ytosine |  | C |  |  |
| **G** | **g**uanine |  |  | G |  |
| **T** | **t**hymine |  |  |  | T |
| **U** | **u**racil |  |  |  | U |
| **W** | **w**eak | A |  |  | T | 2 |
| **S** | **s**trong |  | C | G |  |
| **M** | a**m**ino | A | C |  |  |
| **K** | **k**eto |  |  | G | T |
| **R** | pu**r**ine | A |  | G |  |
| **Y** | p**y**rimidine |  | C |  | T |
| **B** | not A |  | C | G | T | 3 |
| **D** | not C | A |  | G | T |
| **H** | not G | A | C |  | T |
| **V** | not T | A | C | G |  |
| **N** | a**n**y base | A | C | G | T | 4 |