Homework #4

Due October 30th, 11:59pm

Please note that the homework submissions requirements are different from previous homeworks – please read the next section carefully.

Special note: Feel free to use either Python or C/C++ to implement the simulations in this homework. Python may be considerable easier.

Each homework submission must include:

- An archive (.zip or .gz) file of the source code containing:
  - The makefile used to compile the code on Monsoon **(5pts – only if using C++)**
  - All .cpp and .h files **(5pts – only if using C++)**
  - A readme.txt file outlining all modules (if any) needed for the execution of the code and the exact command lines needed to answer homework's questions **(5pts or <u>10pts</u> if not using C++)**
- A full write-up (.pdf of .doc) file containing answers to homework's questions **(5pts or <u>10pts</u> if not using C++)** – screenshots of code output are ok.

The source code must follow the following guidelines:

- No external libraries that implement data structures discussed in class are allowed, unless specifically stated as part of the problem definition. Standard input/output and utilities libraries (e.g. math.h) are ok.
- All external data sources (e.g. input data) must be passed in as a command line argument (no hardcoded paths within the source code.
- Solutions to sub-problems must be executable separately from each other. For example, via a special flag passed as command line argument **(5pts)**

Problem #1 (of 2): Basic simulations

Assume p = 0.005 is the probability of success

A. What is the average number of failures until the first success? Provide estimate of mean and the code to obtain these results.
B. What is the average number of successes in 10,000 trials? Provide estimate of mean and the code to obtain these results.

Problem #2 (of 2): Impact of sequencing errors on problem complexity

Download the genome of a Dengue virus (DENV2) at: http://www.ncbi.nlm.nih.gov/nuccore/NC_001474.2 . OR use the FASTA format sequence at the end of this homework assignment.

A. Generate 100,000 random sequence fragments, each 16 nucleotides long (16-mers) from this virus. Assume uniform random distribution of fragments across the genomic sequence of DENV2.
   - On average, how many **unique** 16-mers have you observed? Use any of the data structures you have used in the previous homeworks (if using C++). If not using C++, built-in datastructures are OK to use here.
   - Does that make sense to you? Explain why or why not.
B. Assume 1% error rate in the sequence fragments – i.e. every base in each fragment has exactly 1% probability to be changed (substitutions only). This means that some fragments may have more than one error. Generate 100,000 sequence fragments with this error rate.
   a. On average, how many **unique** 16-mers have you observed?
   b. Does that make sense to you? Explain why or why not.
C. Repeat 2B with error rate of 5%.
   a. On average, how many unique 16-mers have you observed?
   b. Does that make sense to you? Explain why or why not.

>gi|158976983|ref|NC_001474.2| Dengue virus 2, complete genome

```
AGTTGTTAGTCTACGTGGACCGACAAAGACAGATTCTTTGAGGGAGCTAAGCTCAACGTAGTTCTAACAG
TTTTTTAATTAGAGAGCAGATCTCTGATGAATAACCAACGGAAAAAGGCGAAAAACACGCCTTTCAATAT
GCTGAAACGCGAGAGAAACCGCGTGTCGACTGTGCAACAGCTGACAAAGAGATTCTCACTTGGAATGCTG
CAGGGACGAGGACCATTAAAACTGTTCATGGCCCTGGTGGCGTTCCTTCGTTTCCTAACAATCCCACCAA
CAGCAGGGATATTGAAGAGATGGGGAACAATTAAAAAATCAAAAGCTATTAATGTTTTGAGAGGGTTCAG
GAAAGAGATTGGAAGGATGCTGAACATCTTGAATAGGAGACGCAGATCTGCAGGCATGATCATTATGCTG
ATTCCAACAGTGATGGCGTTCCATTTAACCACACGTAACGGAGAACCACACATGATCGTCAGCAGACAAG
AGAAAGGGAAAAGTCTTCTGTTTAAAACAGAGGATGGCGTGAACATGTGTACCCTCATGGCCATGGACCT
TGGTGAATTGTGTGAAGACACAATCACGTACAAGTGTCCCCTTCTCAGGCAGAATGAGCCAGAAGACATA
GACTGTTGGTGCAACTCTACGTCCACGTGGGTAACTTATGGGACGTGTACCACCATGGGAGAACATAGAA
GAGAAAAAAGATCAGTGGCACTCGTTCCACATGTGGGAATGGGACTGGAGACACGAACTGAAACATGGAT
GTCATCAGAAGGGGCCTGGAAACATGTCCAGAGAATTGAAACTTGGATCTTGAGACATCCAGGCTTCACC
ATGATGGCAGCAATCCTGGCATACACCATAGGAACGACACATTTCCAAAGAGCCCTGATTTTCATCTTAC
TGACAGCTGTCACTCCTTCAATGACAATGCGTTGCATAGGAATGTCAAATAGAGACTTTGTGGAAGGGGT
TTCAGGAGGAAGCTGGGTTGACATAGTCTTAGAACATGGAAGCTGTGTGACGACGATGGCAAAAAACAAA
CCAACATTGGATTTTGAACTGATAAAAACAGAAGCCAAACAGCCTGCCACCCTAAGGAAGTACTGTATAG
AGGCAAAGCTAACCAACACAACAACAGAATCTCGCTGCCCAACACAAGGGGAACCCAGCCTAAATGAAGA
GCAGGACAAAAGGTTCGTCTGCAAACACTCCATGGTAGACAGAGGATGGGGAAATGGATGTGGACTATTT
GGAAAGGGAGGCATTGTGACCTGTGCTATGTTCAGATGCAAAAAGAACATGGAAGGAAAAGTTGTGCAAC
CAGAAAACTTGGAATACACCATTGTGATAACACCTCACTCAGGGGAAGAGCATGCAGTCGGAAATGACAC
AGGAAAACATGGCAAGGAAATCAAAATAACACCACAGAGTTCCATCACAGAAGCAGAATTGACAGGTTAT
GGCACTGTCACAATGGAGTGCTCTCCAAGAACGGGCCTCGACTTCAATGAGATGGTGTTGCTGCAGATGG
AAAATAAAGCTTGGCTGGTGCACAGGCAATGGTTCCTAGACCTGCCGTTACCATGGTTGCCCGGAGCGGA
CACACAAGGGTCAAATTGGATACAGAAAGAGACATTGGTCACTTTCAAAAATCCCCATGCGAAGAAACAG
GATGTTGTTGTTTTAGGATCCCAAGAAGGGGCCATGCACACAGCACTTACAGGGGCCACAGAAATCCAAA
TGTCATCAGGAAACTTACTCTTCACAGGACATCTCAAGTGCAGGCTGAGAATGGACAAGCTACAGCTCAA
AGGAATGTCATACTCTATGTGCACAGGAAAGTTTAAAGTTGTGAAGGAAATAGCAGAAACACAACATGGA
ACAATAGTTATCAGAGTGCAATATGAAGGGGACGGCTCTCCATGCAAGATCCCTTTTGAGATAATGGATT
TGGAAAAAAGACATGTCTTAGGTCGCCTGATTACAGTCAACCCAATTGTGACAGAAAAAGATAGCCCAGT
CAACATAGAAGCAGAACCTCCATTCGGAGACAGCTACATCATCATAGGAGTAGAGCCGGGACAACTGAAG
CTCAACTGGTTTAAGAAAGGAAGTTCTATCGGCCAAATGTTTGAGACAACAATGAGGGGGGCGAAGAGAA
TGGCCATTTTAGGTGACACAGCCTGGGATTTTGGATCCTTGGGAGGAGTGTTTACATCTATAGGAAAGGC
TCTCCACCAAGTCTTTGGAGCAATCTATGGAGCTGCCTTCAGTGGGGTTTCATGGACTATGAAAATCCTC
ATAGGAGTCATTATCACATGGATAGGAATGAATTCACGCAGCACCTCACTGTCTGTGACACTAGTATTGG
TGGGAATTGTGACACTGTATTTGGGAGTCATGGTGCAGGCCGATAGTGGTTGCGTTGTGAGCTGGAAAAA
CAAAGAACTGAAATGTGGCAGTGGGATTTTCATCACAGACAACGTGCACACATGGACAGAACAATACAAG
TTCCAACCAGAATCCCCTTCAAAACTAGCTTCAGCTATCCAGAAAGCCCATGAAGAGGGCATTTGTGGAA
TCCGCTCAGTAACAAGACTGGAGAATCTGATGTGGAAACAAATAACACCAGAATTGAATCACATTCTATC
AGAAAATGAGGTGAAGTTAACTATTATGACAGGAGACATCAAAGGAATCATGCAGGCAGGAAAACGATCT
CTGCGGCCTCAGCCCACTGAGCTGAAGTATTCATGGAAAACATGGGGCAAAGCAAAAATGCTCTCTACAG
AGTCTCATAACCAGACCTTTCTCATTGATGGCCCCGAAACAGCAGAATGCCCCAACACAAATAGAGCTTG
GAATTCGTTGGAAGTTGAAGACTATGGCTTTGGAGTATTCACCACCAATATATGGCTAAAATTGAAAGAA
AAACAGGATGTATTCTGCGACTCAAAACTCATGTCAGCGGCCATAAAAGACAACAGAGCCGTCCATGCCG
ATATGGGTTATTGGATAGAAAGTGCACTCAATGACACATGGAAGATAGAGAAAGCCTCTTTCATTGAAGT
TAAAAACTGCCACTGGCCAAAATCACACACCCTCTGGAGCAATGGAGTGCTAGAAAGTGAGATGATAATT
CCAAAGAATCTCGCTGGACCAGTGTCTCAACACAACTATAGACCAGGCTACCATACACAAATAACAGGAC
CATGGCATCTAGGTAAGCTTGAGATGGACTTTGATTTCTGTGATGGAACAACAGTGGTAGTGACTGAGGA
CTGCGGAAATAGAGGACCCTCTTTGAGAACAACCACTGCCTCTGGAAAACTCATAACAGAATGGTGCTGC
CGATCTTGCACATTACCACCGCTAAGATACAGAGGTGAGGATGGGTGCTGGTACGGGATGGAAATCAGAC
CATTGAAGGAGAAAGAAGAGAATTTGGTCAACTCCTTGGTCACAGCTGGACATGGGCAGGTCGACAACTT
TTCACTAGGAGTCTTGGGAATGGCATTGTTCCTGGAGGAAATGCTTAGGACCCGAGTAGGAACGAAACAT
GCAATACTACTAGTTGCAGTTTCTTTTGTGACATTGATCACAGGGAACATGTCCTTTAGAGACCTGGGAA
GAGTGATGGTTATGGTAGGCGCCACTATGACGGATGACATAGGTATGGGCGTGACTTATCTTGCCCTACT
AGCAGCCTTCAAAGTCAGACCAACTTTTGCAGCTGGACTACTCTTGAGAAAGCTGACCTCCAAGGAATTG
ATGATGACTACTATAGGAATTGTACTCCTCTCCCAGAGCACCATACCAGAGACCATTCTTGAGTTGACTG
ATGCGTTAGCCTTAGGCATGATGGTCCTCAAAATGGTGAGAAATATGGAAAAGTATCAATTGGCAGTGAC
TATCATGGCTATCTTGTGCGTCCCAAACGCAGTGATATTACAAAACGCATGGAAAGTGAGTTGCACAATA
TTGGCAGTGGTGTCCGTTTCCCCACTGCTCTTAACATCCTCACAGCAAAAAACAGATTGGATACCATTAG
CATTGACGATCAAAGGTCTCAATCCAACAGCTATTTTTCTAACAACCCTCTCAAGAACCAGCAAGAAAAG
GAGCTGGCCATTAAATGAGGCTATCATGGCAGTCGGGATGGTGAGCATTTTAGCCAGTTCTCTCCTAAAA
AATGATATTCCCATGACAGGACCATTAGTGGCTGGAGGGCTCCTCACTGTGTGCTACGTGCTCACTGGAC
GATCGGCCGATTTGGAACTGGAGAGAGCAGCCGATGTCAAATGGGAAGACCAGGCAGAGATATCAGGAAG
CAGTCCAATCCTGTCAATAACAATATCAGAAGATGGTAGCATGTCGATAAAAAATGAAGAGGAAGAACAA
ACACTGACCATACTCATTAGAACAGGATTGCTGGTGATCTCAGGACTTTTTCCTGTATCAATACCAATCA
CGGCAGCAGCATGGTACCTGTGGGAAGTGAAGAAACAACGGGCCGGAGTATTGTGGGATGTTCCTTCACC
```

```
CCCACCCATGGGAAAGGCTGAACTGGAAGATGGAGCCTATAGAATTAAGCAAAAAGGGATTCTTGGATAT
TCCCAGATCGGAGCCGGAGTTTACAAAGAAGGAACATTCCATACAATGTGGCATGTCACACGTGGCGCTG
TTCTAATGCATAAAGGAAAGAGGATTGAACCATCATGGGCGGACGTCAAGAAAGACCTAATATCATATGG
AGGAGGCTGGAAGTTAGAAGGAGAATGGAAGGAAGGAGAAGAAGTCCAGGTATTGGCACTGGAGCCTGGA
AAAAATCCAAGAGCCGTCCAAACGAAACCTGGTCTTTTCAAAACCAACGCCGGAACAATAGGTGCTGTAT
CTCTGGACTTTTCTCCTGGAACGTCAGGATCTCCAATTATCGACAAAAAAGGAAAAGTTGTGGGTCTTTA
TGGTAATGGTGTTGTTACAAGGAGTGGAGCATATGTGAGTGCTATAGCCCAGACTGAAAAAAGCATTGAA
GACAACCCAGAGATCGAAGATGACATTTTCCGAAAGAGAAGACTGACCATCATGGACCTCCACCCAGGAG
CGGGAAAGACGAAGAGATACCTTCCGGCCATAGTCAGAGAAGCTATAAAACGGGGTTTGAGAACATTAAT
CTTGGCCCCCACTAGAGTTGTGGCAGCTGAAATGGAGGAAGCCCTTAGAGGACTTCCAATAAGATACCAG
ACCCCAGCCATCAGAGCTGAGCACACCGGGCGGGAGATTGTGGACCTAATGTGTCATGCCACATTTACCA
TGAGGCTGCTATCACCAGTTAGAGTGCCAAACTACAACCTGATTATCATGGACGAAGCCCATTTCACAGA
CCCAGCAAGTATAGCAGCTAGAGGATACATCTCAACTCGAGTGGAGATGGGTGAGGCAGCTGGGATTTTT
ATGACAGCCACTCCCCCGGGAAGCAGAGACCCATTTCCTCAGAGCAATGCACCAATCATAGATGAAGAAA
GAGAAATCCCTGAACGTTCGTGGAATTCCGGACATGAATGGGTCACGGATTTTAAAGGGAAGACTGTTTG
GTTCGTTCCAAGTATAAAAGCAGGAAATGATATAGCAGCTTGCCTGAGGAAAAATGGAAAGAAAGTGATA
CAACTCAGTAGGAAGACCTTTGATTCTGAGTATGTCAAGACTAGAACCAATGATTGGGACTTCGTGGTTA
CAACTGACATTTCAGAAATGGGTGCCAATTTCAAGGCTGAGAGGGTTATAGACCCCAGACGCTGCATGAA
ACCAGTCATACTAACAGATGGTGAAGAGCGGGTGATTCTGGCAGGACCTATGCCAGTGACCCACTCTAGT
GCAGCACAAAGAAGAGGGAGAATAGGAAGAAATCCAAAAAATGAGAATGACCAGTACATATACATGGGGG
AACCTCTGGAAAATGATGAAGACTGTGCACACTGGAAAGAAGCTAAAATGCTCCTAGATAACATCAACAC
GCCAGAAGGAATCATTCCTAGCATGTTCGAACCAGAGCGTGAAAAGGTGGATGCCATTGATGGCGAATAC
CGCTTGAGAGGAGAAGCAAGGAAAACCTTTGTAGACTTAATGAGAAGAGGAGACCTACCAGTCTGGTTGG
CCTACAGAGTGGCAGCTGAAGGCATCAACTACGCAGACAGAAGGTGGTGTTTTGATGGAGTCAAGAACAA
CCAAATCCTAGAAGAAAACGTGGAAGTTGAAATCTGGACAAAAGAAGGGGAAAGGAAGAAATTGAAACCC
AGATGGTTGGATGCTAGGATCTATTCTGACCCACTGGCGCTAAAAGAATTTAAGGAATTTGCAGCCGGAA
GAAAGTCTCTGACCCTGAACCTAATCACAGAAATGGGTAGGCTCCCAACCTTCATGACTCAGAAGGCAAG
AGACGCACTGGACAACTTAGCAGTGCTGCACACGGCTGAGGCAGGTGGAAGGGCGTACAACCATGCTCTC
AGTGAACTGCCGGAGACCCTGGAGACATTGCTTTTACTGACACTTCTGGCTACAGTCACGGGAGGGATCT
TTTTATTCTTGATGAGCGGAAGGGGCATAGGGAAGATGACCCTGGGAATGTGCTGCATAATCACGGCTAG
CATCCTCCTATGGTACGCACAAATACAGCCACACTGGATAGCAGCTTCAATAATACTGGAGTTTTTTCTC
ATAGTTTTGCTTATTCCAGAACCTGAAAAACAGAGAACACCCCAAGACAACCAACTGACCTACGTTGTCA
TAGCCATCCTCACAGTGGTGGCCGCAACCATGGCAAACGAGATGGGTTTCCTAGAAAAAACGAAGAAAGA
TCTCGGATTGGGAAGCATTGCAACCCAGCAACCCGAGAGCAACATCCTGGACATAGATCTACGTCCTGCA
TCAGCATGGACGCTGTATGCCGTGGCCACAACATTTGTTACACCAATGTTGAGACATAGCATTGAAAATT
CCTCAGTGAATGTGTCCCTAACAGCTATAGCCAACCAAGCCACAGTGTTAATGGGTCTCGGGAAAGGATG
GCCATTGTCAAAGATGGACATCGGAGTTCCCCTTCTCGCCATTGGATGCTACTCACAAGTCAACCCCATA
ACTCTCACAGCAGCTCTTTTCTTATTGGTAGCACATTATGCCATCATAGGGCCAGGACTCCAAGCAAAAG
CAACCAGAGAAGCTCAGAAAAGAGCAGCGGCGGGCATCATGAAAAACCCAACTGTCGATGGAATAACAGT
GATTGACCTAGATCCAATACCTTATGATCCAAAGTTTGAAAAGCAGTTGGGACAAGTAATGCTCCTAGTC
CTCTGCGTGACTCAAGTATTGATGATGAGGACTACATGGGCTCTGTGTGAGGCTTTAACCTTAGCTACCG
GGCCCATCTCCACATTGTGGGAAGGAAATCCAGGGAGGTTTTGGAACACTACCATTGCGGTGTCAATGGC
TAACATTTTTAGAGGGAGTTACTTGGCCGGAGCTGGACTTCTCTTTTCTATTATGAAGAACACAACCAAC
ACAAGAAGGGGAACTGGCAACATAGGAGAGACGCTTGGAGAGAAATGGAAAAGCCGATTGAACGCATTGG
GAAAAAGTGAATTCCAGATCTACAAGAAAAGTGGAATCCAGGAAGTGGATAGAACCTTAGCAAAAGAAGG
CATTAAAAGAGGAGAAACGGACCATCACGCTGTGTCGCGAGGCTCAGCAAAACTGAGATGGTTCGTTGAG
AGAAACATGGTCACACCAGAAGGGAAAGTAGTGGACCTCGGTTGTGGCAGAGGAGGCTGGTCATACTATT
GTGGAGGACTAAAGAATGTAAGAGAAGTCAAAGGCCTAACAAAAGGAGGACCAGGACACGAAGAACCCAT
CCCCATGTCAACATATGGGTGGAATCTAGTGCGTCTTCAAAGTGGAGTTGACGTTTTCTTCATCCCGCCA
GAAAAGTGTGACACATTATTGTGTGACATAGGGGAGTCATCACCAAATCCCACAGTGGAAGCAGGACGAA
CACTCAGAGTCCTTAACTTAGTAGAAAATTGGTTGAACAACAACACTCAATTTTGCATAAAGGTTCTCAA
CCCATATATGCCCTCAGTCATAGAAAAAATGGAAGCACTACAAAGGAAATATGGAGGAGCCTTAGTGAGG
AATCCACTCTCACGAAACTCCACACATGAGATGTACTGGGTATCCAATGCTTCCGGGAACATAGTGTCAT
CAGTGAACATGATTTCAAGGATGTTGATCAACAGATTTACAATGAGATACAAGAAAGCCACTTACGAGCC
GGATGTTGACCTCGGAAGCGGAACCCGTAACATCGGGATTGAAAGTGAGATACCAAACCTAGATATAATT
GGGAAAAGAATAGAAAAAATAAAGCAAGAGCATGAAACATCATGGCACTATGACCAAGACCACCCATACA
AAACGTGGGCATACCATGGTAGCTATGAAACAAAACAGACTGGATCAGCATCATCCATGGTCAACGGAGT
GGTCAGGCTGCTGACAAAACCTTGGGACGTCGTCCCCATGGTGACACAGATGGCAATGACAGACACGACT
CCATTTGGACAACAGCGCGTTTTTAAAGAGAAAGTGGACACGAGAACCCAAGAACCGAAAGAAGGCACGA
AGAAACTAATGAAAATAACAGCAGAGTGGCTTTGGAAAGAATTAGGGAAGAAAAAGACACCCAGGATGTG
CACCAGAGAAGAATTCACAAGAAAGGTGAGAAGCAATGCAGCCTTGGGGGCCATATTCACTGATGAGAAC
AAGTGGAAGTCGGCACGTGAGGCTGTTGAAGATAGTAGGTTTTGGGAGCTGGTTGACAAGGAAAGGAATC
TCCATCTTGAAGGAAAGTGTGAAACATGTGTGTACAACATGATGGGAAAAAGAGAGAAGAAGCTAGGGGA
ATTCGGCAAGGCAAAAGGCAGCAGAGCCATATGGTACATGTGGCTTGGAGCACGCTTCTTAGAGTTTGAA
GCCCTAGGATTCTTAAATGAAGATCACTGGTTCTCCAGAGAGAACTCCCTGAGTGGAGTGGAAGGAGAAG
GGCTGCACAAGCTAGGTTACATTCTAAGAGACGTGAGCAAGAAAGAGGGAGGAGCAATGTATGCCGATGA
```

```
CACCGCAGGATGGGATACAAGAATCACACTAGAAGACCTAAAAAATGAAGAAATGGTAACAAACCACATG
GAAGGAGAACACAAGAAACTAGCCGAGGCCATTTTCAAACTAACGTACCAAAACAAGGTGGTGCGTGTGC
AAAGACCAACACCAAGAGGCACAGTAATGGACATCATATCGAGAAGAGACCAAAGAGGTAGTGGACAAGT
TGGCACCTATGGACTCAATACTTTCACCAATATGGAAGCCCAACTAATCAGACAGATGGAGGGAGAAGGA
GTCTTTAAAAGCATTCAGCACCTAACAATCACAGAAGAAATCGCTGTGCAAAACTGGTTAGCAAGAGTGG
GGCGCGAAAGGTTATCAAGAATGGCCATCAGTGGAGATGATTGTGTTGTGAAACCTTTAGATGACAGGTT
CGCAAGCGCTTTAACAGCTCTAAATGACATGGGAAAGATTAGGAAAGACATACAACAATGGGAACCTTCA
AGAGGATGGAATGATTGGACACAAGTGCCCTTCTGTTCACACCATTTCCATGAGTTAATCATGAAAGACG
GTCGCGTACTCGTTGTTCCATGTAGAAACCAAGATGAACTGATTGGCAGAGCCCGAATCTCCCAAGGAGC
AGGGTGGTCTTTGCGGGAGACGGCCTGTTTGGGGAAGTCTTACGCCCAAATGTGGAGCTTGATGTACTTC
CACAGACGCGACCTCAGGCTGGCGGCAAATGCTATTTGCTCGGCAGTACCATCACATTGGGTTCCAACAA
GTCGAACAACCTGGTCCATACATGCTAAACATGAATGGATGACAACGGAAGACATGCTGACAGTCTGGAA
CAGGGTGTGGATTCAAGAAAACCCATGGATGGAAGACAAAACTCCAGTGGAATCATGGGAGGAAATCCCA
TACTTGGGGAAAAGAGAAGACCAATGGTGCGGCTCATTGATTGGGTTAACAAGCAGGGCCACCTGGGCAA
AGAACATCCAAGCAGCAATAAATCAAGTTAGATCCCTTATAGGCAATGAAGAATACACAGATTACATGCC
ATCCATGAAAAGATTCAGAAGAGAAGAGGAAGAAGCAGGAGTTCTGTGGTAGAAAGCAAAACTAACATGA
AACAAGGCTAGAAGTCAGGTCGGATTAAGCCATAGTACGGAAAAAACTATGCTACCTGTGAGCCCCGTCC
AAGGACGTTAAAAGAAGTCAGGCCATCATAAATGCCATAGCTTGAGTAAACTATGCAGCCTGTAGCTCCA
CCTGAGAAGGTGTAAAAAATCCGGGAGGCCACAAACCATGGAAGCTGTACGCATGGCGTAGTGGACTAGC
GGTTAGAGGAGACCCCTCCCTTACAAATCGCAGCAACAATGGGGGCCCAAGGCGAGATGAAGCTGTAGTC
TCGCTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCGAAACAAAAAACAGCATATTGACGCTGGGAAAGA
CCAGAGATCCTGCTGTCTCCTCAGCATCATTCCAGGCACAGAACGCCAGAAAATGGAATGGTGCTGTTGA
ATCAACAGGTTCT
```