

# Homework #4

Due April 1<sup>st</sup>, 11:59pm

---

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)**
  - All .cpp and .h files **(5pts)**
  - 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)**
- A full write-up (.pdf or .doc) file containing answers to homework's questions **(5pts)** – 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)**
- 

## Special notes:

- on this homework you can use C++ default random number generator (srand) or download whatever random number generator you wish.
- If you are unable to get the Smith-Waterman algorithm to work in problem #1, you are free to download an external library that implements this algorithm for problem #2 (point that out in your write-up so that the grader knows)

### Problem #1 (of 2): Implement Smith-Waterman algorithm

- A. Write a function in C++ that would implement the Smith-Waterman alignment between two genomic sequences.
- The function must take a genomic similarity scoring matrix (+2 for match, -1 for mismatch) and a gap penalty (-3) as an input.
  - The function must return:
    - score for best alignment
    - alignment in text format (hint: use a struct of 3 character arrays, 2 for sequences one for alignment codes {x, l, } – use whitespace for gaps). See below.
  - **Test your code** using the SARS-COV2 viral genome found in appendix A at the bottom of this homework assignment and sequence fragments found in appendix B at the bottom of this homework.
  - You will need to submit a screenshot of the output as part of the homework write up (it should look something like the text below)

```
GATTA
| x | |
G _CTA
```

- B. Generate 1K, 10K, 100K, and 1M (million) **completely random** genomic sequences (50nt) to use as targets for alignment and use SARS-COV2 genome as subject. Perform alignment of the queries to the subject sequence and record time to completion (in seconds / minutes).

**Special note:** Depending on the speed of your alignment implementation, this homework may take hours or days to complete. The goal is to get a sense for how slow these 'optimal' alignment algorithms are... for the explicit purpose of establishing a baseline to be able to compare the improved algorithms and data structures we will be discussing later in the course. Please be aware of this and 'give up' on larger benchmarks as appropriate.

## **Problem #2 (of 2): Having a BLAST**

- A. **Implement a seed-based Smith Waterman.** This means:
- Use the genome in Appendix A and break it down into seeds with **word size = 11**.
  - Load your seeds (created in part A) into memory
  - For each read disassemble it into k-mers (of size 11). Compare your read k-mers to the SARS-COV2 seeds.
  - If a seed-match is found, extend the seed by cutting out the appropriate segment of the subject (Genome of SARS-COV2) and running the Smith Waterman on the two sequences (original read and the segment from SARS-COV2)
    - Beware of edge cases
    - **Ok to just expand one seed and be done** (multiple seeds from a read can be found, typically necessitating multiple seed expansions and decision on what is the best alignment)
  - **Test your code** on the 50-mers I have provided below in Appendix B. You must report alignment for the 50-mers I've provided as part of the homework solution.
- B. **Test your code** on a set of 1K, 10K, 100K, and 1M (million) **completely random 50-mers**, aligning them to SARS-COV2 genome. How long did it take? Compare it to the results in problem 1B.
- C. **Test algorithm's exhaustiveness.** Randomly select 100,000 fragments from the SARS-COV2 genome and use these fragments to query the SARS-COV2 genome using the seed-based SW you implemented in part A. **How many fragments were you able to find?** Now introduce random errors into your 100,000 fragments at a 5% per-base error rate (every character has a 5% change of being changed to some other random character). Use these error-filled 100,000 fragments to query the SARS-COV2 genome again. **How many fragments were you able to find?**

## Appendix A: SARS-COV2 Genome

>NC\_045512.2\_SARS\_COV2\_complete\_genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTTCGATCTCTTGTAGATCTGTTCTCTAAA  
CGAACTTTAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC  
TAATTACTGTCGTTGACAGGACACGAGTAACCTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG  
TTGCAGCCGATCATCAGCACATCTAGGTTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC  
CCTGGTTTCAACGAGAAAAACACACGTCCAACCTCAGTTTGCCTGTTTTACAGGTTTCGCGACGTGCTCGTAC  
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG  
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTTCATCAAACGTTCCGAT  
GCTCGAAGTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAAGTGAAGGCATTAGTACGGTC  
GTAGTGGTGAGACACTTGGTGTCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT  
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA  
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTTCAAGAAAACCTGGAACACTAAACATAGCAGTGGTG  
TTACCCGTGAACCTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACCTTCTGTGG  
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG  
TCCGAACAACCTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG  
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA  
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA  
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC  
CAAATGAATGCAACCAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA  
GACGGGCGATTTTGTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACT  
ACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAG  
GACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAACACCATTCTTCGTAAGGGTGGTCG  
CACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCA  
CGTGCTAGCGCTAACATAGGTTGTAACCATAACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACA  
ACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAACTTAATGAAGA  
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAACTGTGAAAGGTTTGGAT  
TATAAAGCATTAACAATTTGTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAGCTAAAAAG  
GTGCCTGGAATATTGGTGAACAGAAATCAATAGTACTGAGTCTCTTTATGCATTTGCATCAGAGGCTGCTCG  
TGTTGTACGATCAATTTTCTCCCGCACTCTTGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCC  
GCTATAACAATACTAGATGGAATTTACAGTATTCACTGAGACTCATTGATGCTATGATGTTTACATCTG  
ATTTGGCTACTAACAATCTAGTTGTAATGGCCTACATTACAGGTGGTGTGTTTCAAGTTGACTTCGCAGTG  
GCTAACTAACATCTTTGGCACTGTTTATGAAAACTCAAACCCGTCCTTGATTGGCTTGAAGAGAAGTTT  
AAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGTTAAATTTATCTCAACCTGTGCTTGTGAAA  
TTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAAGGAGAGTGTTTCAAGACATTCTTTAAGCTTGT  
AAATAAATTTTGGCTTTGTGTGCTGACTCTATCATTATTGGTGGAGCTAACTTAAAGCCTTGAATTTA  
GGTGAACATTTGTACGCACTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAACTGGCC  
TACTCATGCCTCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCACAGAAGTGTT  
AACAGAGGAAGTTGTCTTGAAAACCTGGTGATTTACAACCATTAGAACAACCTACTAGTGAAGCTGTTGAA  
GCTCCATTGGTTGGTACACCAAGTTTGTATTAACGGGCTTATGTTGCTCGAAATCAAAGACACAGAAAAGT  
ACTGTGCCCTTGCACCTAATATGATGGTAACAAACAATACCTTCACACTCAAAGGCGGTGCACCAACAAA  
GGTTACTTTTGGTGATGACACTGTGATAGAAGTGCAAGGTTACAAGAGTGTAATATCACTTTTGAACCTT  
GATGAAAGGATTGATAAAGTACTTAATGAGAAGTGCTCTGCCTATACAGTTGAACTCGGTACAGAAGTAA  
ATGAGTTTCGCTGTGTTGTGGCAGATGCTGTCTATAAAAACTTTGCAACCAGTATCTGAATTACTTACACC  
ACTGGGCATTGATTTAGATGAGTGGAGTATGGCTACATACTACTTATTTGATGAGTCTGGTGAGTTTAA  
TTGGCTTACATATGTATTGTTCTTTCTACCTCCAGATGAGGATGAAGAAGAAGGTGATTGTGAAGAAG  
AAGAGTTTGAAGCATCAACTCAATATGAGTATGGTACTGAAGATGATTACCAAGGTAAACCTTTGGAATT  
TGGTGCCACTTCTGCTGCTCTTCAACCTGAAGAAGAGCAAGAAGAAGATTGGTTAGATGATGATAGTCAA  
CAAACCTGTTGGTCAACAAGACGGCAGTGAGGACAATCAGACAACCTACTATTCAAACAATTGTTGAGGTTT  
AACCTCAATTAGAGATGGAACCTTACACCAAGTTGTTTCAAGTATTGAAGTGAATAGTTTTAGTGGTTATTT  
AAAACCTTACTGACAATGTATACATTAAAAATGCAGACATTGTGGAAGAAGCTAAAAAGGTAAACCAACA  
GTGGTTGTTAATGCAGCCAATGTTTACCTTAAACATGGAGGAGGTGTTGCAGGAGCCTTAAATAAGGCTA  
CTAACAATGCCATGCAAGTTGAATCTGATGATTACATAGCTACTAATGGACCACTTAAAGTGGGTGGTAG  
TTGTGTTTTAAGCGGACACAATCTTGCTAAACACTGTCTTCATGTTGTGCGCCCAAATGTTAACAAGGT  
GAAGACATTCAACTTCTTAAGAGTGCTTATGAAAATTTTAAATCAGCACGAAGTTCTACTTGCACCATAT  
TATCAGCTGGTATTTTTGGTGCTGACCCTATACATTCTTTAAGAGTTTGTGTAGATACTGTTTCGACAAA

[illegible]

TCTTTACTTTGATAAAGCTGGTCAAAAGACTTATGAAAGACATTCTCTCTCTCATTGTTAACTTAGAC  
AACCTGAGAGCTAATAACACTAAAGGTTCAATTGCCTATTAATGTTATAGTTTTTGTATGGTAAATCAAAAT  
GTGAAGAATCATCTGCAAAATCAGCGTCTGTTTACTACAGTCAGCTTATGTGTCAACCTATACTGTTACT  
AGATCAGGCATTAGTGTCTGATGTTGGTGATAGTGCAGGAAAGTTGCAGTTAAAATGTTTGATGCTTACGTT  
AATACGTTTTTCATCAACTTTTAAACGTACCAATGGAAAACTCAAAACACTAGTTGCAACTGCAGAAGCTG  
AACTTGCAAAGAATGTGTCCTTAGACAATGTCTTATCTACTTTTTATTTTCAGCAGCTCGGCAAGGGTTTTGT  
TGATTGAGATGTAGAACTAAAGATGTTGTTGAATGTCTTAAATTGTCACATCAATCTGACATAGAAGTT  
ACTGGCGATAGTTGTAATAACTATATGCTCACCTATAACAAAGTTGAAAACATGACACCCCGTGACCTTG  
GTGCTTGTATTGACTGTAGTGCGCGTCATATTAATGCGCAGGTAGCAAAAAGTCACAACATTGCTTTGAT  
ATGGAACGTTAAAGATTTTCATGTCAATTGTCTGAACAACTACGAAAACAAATACGTAGTGCTGTCAAAAAG  
AATAACTTACCTTTTTAAGTTGACATGTGCAACTACTAGACAAGTTGTTAATGTTGTAACAACAAGATAG  
CACTTAAGGGTGGTAAAATTGTTAATAATTGGTTGAAGCAGTTAATTAAGTTACACTTGTGTTCCTTTTT  
TGTTGCTGCTATTTTTCTATTTAATAACACCTGTTTCATGTCTGCTAAACATACTGACTTTTTCAAGTGAA  
ATCATAGGATACAAGGCTATTGATGGTGGTGTCACTCGTGACATAGCATCTACAGATACTTGTTTTTGCTA  
ACAAACATGCTGATTTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATACTAATGACAAAGCTTGCCC  
ATTGATTGCTGCAGTCATAACAAGAGAAGTGGGTTTTGTGCTGCCTGGTTTTGCCTGGCAGCATATTACGC  
ACAATAATGGTGACTTTTTGTCATTTCTTACCTAGAGTTTTTAGTGAGTTGGTAACATCTGTTACACAC  
CATCAAACTTATAGAGTACACTGACTTTGCAACATCAGCTTGTGTTTTGGCTGCTGAATGTACAATTTT  
TAAAGATGCTTCTGGTAAGCCAGTACCATATTGTTATGATAACCAATGTACTAGAAGGTTCTGTTGCTTAT  
GAAAGTTTACGCCCTGACACACGTTATGTGCTCATGGATGGCTCTATTATTCAATTTCCCTAACACCTACC  
TTGAAGGTTCTGTTAGAGTGGTAACAACCTTTTGATTCTGAGTACTGTAGGCACGGCACTTGTGAAAGATC  
AGAAGCTGGTGGTTGTGTATCTACTAGTGGTAGATGGGTACTTAACAATGATTATTACAGATCTTTACCA  
GGAGTTTTCTGTGGTGTAGATGCTGTAAATTTACTTACTAATATGTTTACACCACTAATTCAACCTATTG  
GTGCTTTGGACATATCAGCATCTATAGTAGCTGGTGGTATTGTAGCTATCGTAGTAACATGCCTTGCCCTA  
CTATTTTATGAGGTTTAGAAGAGCTTTTGGTGAATACAGTCATGTAGTTGCCTTTAATACTTTACTATTC  
CTTATGTCATTCACTGTACTCTGTTTAAACACAGTTTACTCATTCTTACCTGGTGGTTTATTCTGTTATTT  
ACTTGTACTTGACATTTTATCTTACTAATGATGTTTCTTTTTTAGCACATATTCACTGGATGGTTATGTT  
CACACCTTTAGTACCTTTCTGGATAACAATTGCTTATATCATTGTTATTTCCACAAAGCATTCTATTGG  
TTCTTTAGTAATTAACCTAAGAGACGTGTAGTCTTTAATGGTGGTTTCTTTTAGTACTTTTGAAGAAGTG  
CGCTGTGCACCTTTTTGTTAAATAAAGAAATGTATCTAAAGTTGCGTAGTGATGTGCTATTACCTCTTAC  
GCAATATAATAGATACTTAGCTCTTTATAATAAGTACAAGTATTTTAGTGGAGCAATGGATACAACCTAGC  
TACAGAGAAGCTGCTTGTGTGTCATCTCGCAAAGGCTCTCAATGACTTCAGTAACCTCAGGTTCTGATGTTT  
TTTACCAACCACCACAAACCTCTATCACCTCAGCTGTTTTGCAGAGTGGTTTTAGAAAAATGGCATTCCC  
ATCTGGTAAAGTTGAGGGTTGTATGGTACAAGTAACCTGTGGTACAACCTACACTTAACGGTCTTTGGCTT  
GATGACGTAGTTTACTGTCCAAGACATGTGATCTGCACCTCTGAAGACATGCTTAACCTAATTATGAAG  
ATTTACTCATTTCGTAAGTCTAATCATAATTTCTTGGTACAGGCTGGTAATGTTCAACTCAGGGTTATTGG  
ACATTCTATGCAAAATTGTGTACTTAAGCTTAAGGTTGATACAGCCAATCCTAAGACACCTAAGTATAAG  
TTTGTTCGCATTCAACCAGGACAGACTTTTTTCACTGTTAGCTTGTACAATGGTTTACCATCTGGTGGTTT  
ACCAATGTGCTATGAGGCCCAATTTCACTATTAAGGGTTTCACTTCTTAATGGTTTCACTGTTGTTGG  
TTTTAACATAGATTATGACTGTGTCTCTTTTTGTTACATGCACCATATGGAATTACCAACTGGAGTTTCA  
GCTGGCACAGACTTAGAAGGTAACCTTTTATGGACCTTTTGTGACAGGCAAACAGCACAAAGCAGCTGGTA  
CGGACACAACCTATTACAGTTAATGTTTTAGCTTGGTTGTACGCTGCTGTTATAAATGGAGACAGGTGGTT  
TCTCAATCGATTTACCACAACCTCTAATGACTTTAACCTTGTGGCTATGAAGTACAATTATGAACCTCTA  
ACACAAGACCATGTTGACATACTAGGACCTCTTTCTGCTCAAACCTGGAATTGCCGTTTTAGATATGTGTG  
CTTCATTAAAAGAATTACTGCAAAATGGTATGAATGGACGTACCATATTGGGTAGTGCTTTATTAGAAGA  
TGAATTTACACCTTTTGATGTTGTTAGACAATGCTCAGGTGTTACTTTCCAAAGTGCAGTGAAAAGAACA  
ATCAAGGGTACACACCCTGGTTGTTACTCACAATTTTGAACCTTCACTTTTAGTTTTAGTCCAGAGTACTC  
AATGGTCTTTGTTCTTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTATGTC  
TGCTTTTGCAATGATGTTTGTCAAACATAAGCATGCATTTCTCTGTTTGTGTTTTGTTACCTTCTCTTGCC  
ACTGTAGCTTATTTAATATGGTCTATATGCCTGCTAGTTGGGTGATGCGTATTATGACATGGTTGGATA  
TGTTTGATACTAGTTTGTCTGGTTTTAAGCTAAAAGACTGTGTTATGTATGCATCAGCTGTAGTGTTACT  
AATCCTTATGACAGCAAGAAGTGTGTATGATGATGGTGCTAGGAGAGTGTGGACACTTATGAATGTCTTG  
ACACTCGTTTTATAAAGTTTATTATGGTAATGCTTTAGATCAAGCCATTTCCATGTGGGCTCTTATAATCT  
CTGTTACTTCTAACTACTCAGGTGTAGTTACAACGTGCATGTTTTTGGCCAGAGGTATTGTTTTTATGTG  
TGTTGAGTATTGCCCTATTTTCTTCATAACTGGTAATACACTTCAGTGTATAATGCTAGTTTTATTGTTTC  
TTAGGCTATTTTTGTACTTGTACTTTGGCCTCTTTTGTGTTACTCAACCGCTACTTTAGACTGACTCTTG  
GTGTTTATGATTACTTAGTTTCTACACAGGAGTTTAGATATATGAATTCACAGGGACTACTCCCACCCAA

GAATAGCATAGATGCCTTCAAACCTCAACATTAATTTGTTGGGTGTTGGTGGCAAACCTTGTATCAAAGTA  
GCCACTGTACAGTCTAAAATGTGAGATGTAAAGTGCACATCAGTAGTCTTACTCTCAGTTTTTGAACAAC  
TCAGAGTAGAATCATCATCTAAATTTGTGGGCTCAATGTGTCCAGTTACACAATGACATTCTCTTAGCTAA  
AGATACTACTGAAGCCTTTGAAAAAATGGTTTTCACTACTTTCTGTTTTGCTTTCCATGCAGGGTGTGTGA  
GACATAAACAAGCTTTGTGAAGAAATGCTGGACAACAGGGCAACCTTACAAGCTATAGCCTCAGAGTTTA  
GTTCCCTTCCATCATATGCAGCTTTTGTACTGCTCAAGAAGCTTATGAGCAGGCTGTTGCTAATGGTGA  
TTCTGAAGTTGTTCTTAAAAAGTTGAAGAAGTCTTTGAATGTGGCTAAATCTGAATTTGACCGTGATGCA  
GCCATGCAACGTAAGTTGGAAAAGATGGCTGATCAAGCTATGACCCAAATGTATAAACAGGCTAGATCTG  
AGGACAAGAGGGCAAAAGTTACTAGTGCTATGCAGACAATGCTTTTCACTATGCTTAGAAAAGTTGGATAA  
TGATGCACTCAACAACATTATCAACAATGCAAGAGATGGTTGTGTTCCCTTGAACATAATACCTCTTACA  
ACAGCAGCCAAACTAATGGTTGTGCATACCAGACTATAACACATATAAAAAATACGTGTGATGGTACAACAT  
TTACTTATGCATCAGCATTGTGGGAAATCCAACAGGTTGTAGATGCAGATAGTAAAATTGTTCAACTTAG  
TGAAATTAGTATGGACAATTCACCTAATTTAGCATGGCCTCTTATTGTAACAGCTTTAAGGGCCAATTCT  
GCTGTCAAATTACAGAATAATGAGCTTAGTCCTGTTGCACTACGACAGATGTCTTGTGCTGCCGGTACTA  
CACAACTGCTTGCCTGATGACAATGCGTTAGCTTACTACAACACAACAAAGGGAGGTAGGTTTGTACT  
TGCCTGTTATCCGATTTACAGGATTTGAAATGGGCTAGATTCCCTAAGAGTGATGGAAGTGGTACTATC  
TATACAGAAGTGAACACCTTGTAGGTTTGTACAGACACACCTAAAGGTCCTAAAGTGAAGTATTTAT  
ACTTTATTAAAGGATTAAACAACCTAAATAGAGGTATGGTACTTGGTAGTTTAGCTGCCACAGTACGTCT  
ACAAGCTGGTAATGCAACAGAAGTGCCTGCCAATTCACTGTATTATCTTTCTGTGCTTTTGTGTAGAT  
GCTGCTAAAGCTTACAAAGATTATCTAGCTAGTGGGGGACAACCAATCACTAATTGTGTTAAGATGTTGT  
GTACACACACTGGTACTGGTCAGGCAATAACAGTTACACCGGAAGCCAATATGGATCAAGAATCCTTTGG  
TGGTGCATCGTGTGTGTCTGTACTGCCGTTGCCACATAGATCATCAAATCCTAAAGGATTTTGTGACTTA  
AAAGGTAAGTATGTACAAATACCTACAACCTTGTGCTAATGACCCTGTGGGTTTTACACTTAAAAACACAG  
TCTGTACCGTCTGCGGTATGTGGAAAGGTTATGGCTGTAGTTGTGATCAACTCCGCGAAGCCATGCTTCA  
GTCAGCTGATGCACAATCGTTTTTAAACGGGTTTGCCTGTAAAGTGCAGCCCGTCTTACACCGTGCAGCA  
CAGGCACTAGTACTGATGTGCTATACAGGGCTTTTGACATCTACAATGATAAAGTAGCTGGTTTTGCTAA  
ATTCCTAAAACTAATTGTTGTGCTTCCAAGAAAAGGACGAAGATGACAATTTAATTGATTCTTACTTT  
GTAGTTAAGAGACACACTTTCTCTAACTACCAACTGAAGAAACAATTTATAATTTACTTAAAGGATTGTC  
CAGCTGTTGCTAAACATGACTTCTTTAAGTTTGAAGATAGACGGTGACATGGTACCACATATATACGTC  
ACGTCTTTACTAAATACACAATGGCAGACCTCGTCTATGCTTTAAGGCATTTTGTGATGAAGTAATTTGTGAC  
ACATTTAAAGAAATACTTGTACATACAATTGTTGTGATGATGATTATTTCAATAAAAAGGACTGGTATG  
ATTTTGTAGAAAACCCAGATATATTACGCGTATACGCCAAGTTAGGTGAACGTGTACGCCAAGCTTTGTT  
AAAAACAGTACAATTCTGTGATGCCATGCGAAATGCTGGTATTGTTGGTGTACTGACATTAGATAATCAA  
GATCTCAATGGTAAGTGGTATGATTTGCGGTGATTTTACATAACACCGCCAGGTAGTGGAGTTCCTGTTG  
TAGATTCTTATTATTATTGTTAATGCCTATATTAACCTTGACCAGGGCTTTAACTGCAGAGTCACATGT  
TGACACTGACTTAAACAAAGCCTTACATTAAGTGGGATTTGTTAAATATGACTTCACGGAAGAGAGGTTA  
AAACTCTTTGACCGTTATTTTAAATATTGGGATCAGACATACCACCCAAATTGTGTTAACTGTTTGGATG  
ACAGATGCATTCTGCATTGTGCAAACCTTAAATGTTTTATTCTCTACAGTGTTCACCTACAAGTTTTGG  
ACCACTAGTGAGAAAAATATTTGTTGATGGTGTTCATTTGTAGTTTCACTGGATACCACTTACAGAGAG  
CTAGGTGTTGTACATAATCAGGATGTAACTTACATAGCTCTAGACTTAGTTTTAAGGAATTACTTGTGT  
ATGCTGCTGACCCTGCTATGCACGCTGCTTCTGGTAATCTATTACTAGATAAACGCACTACGTGCTTTTC  
AGTAGCTGCACTTACTAACAATGTTGCTTTTCAAACCTGTCAAACCCGGTAATTTTAAACAAAGACTTCTAT  
GACTTTGCTGTGTCTAAGGGTTTCTTTAAGGAAGGAAGTTCTGTTGAATTTAAACACTTCTTCTTTGCTC  
AGGATGGTAATGCTGCTATCAGCGATTATGACTACTATCGTTATAATCTACCAACAATGTGTGATATCAG  
ACAACACTATTTGTAGTTGAAGTTGTTGATAAGTACTTTGATTGTTACGATGGTGGCTGTATTAATGCT  
AACCAAGTCATCGTCAACAACCTAGACAAATCAGCTGGTTTTCCATTTAATAAATGGGGTAAGGCTAGAC  
TTTATTATGATTCAATGAGTTATGAGGATCAAGATGCATTTTCGCATATACAAAACGTAATGTCTATCCC  
TACTATAACTCAAATGAATCTTAAAGTATGCCATTAGTGCAAGAATAGAGCTCGCACCGTAGCTGGTGTCT  
TCTATCTGTAGTACTATGACCAATAGACAGTTTCATCAAAAATTATTGAAATCAATAGCCGCCACTAGAG  
GAGCTACTGTAGTAATTGGAACAAGCAAATCTATGGTGGTTGGCACAACATGTTAAAACTGTTTATAG  
TGATGTAGAAAACCTCACCTTATGGGTTGGGATTATCCTAAATGTGATAGAGCCATGCCTAACATGCTT  
AGAATTATGGCCTCACTTGTCTTGTCTCGCAAACATACAACGTGTTGTAGCTTGTACACCGTTTCTATA  
GATTAGCTAATGAGTGTGCTCAAGTATTGAGTGAAATGGTCATGTGTGGCGGTTCACTATATGTTAAACC  
AGGTGGAACCTCATCAGGAGATGCCACAACCTGCTTATGCTAATAGTGTTTTTAACATTTGTCAAGCTGTC  
ACGGCCAATGTTAATGCATTTTATCTACTGATGGTAACAAAATTGCCGATAAGTATGTCCGCAATTTAC  
AACACAGACTTTTATGAGTGTCTCTATAGAAATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGC  
ATATTTGCGTAACATTTCTCAATGATGATACTCTCTGACGATGCTGTTGTGTGTTTCAATAGCACTTAT

GCATCTCAAGGTCTAGTGGCTAGCATAAAGAACTTTAAGTCAGTTCTTTATTATCAAAACAATGTTTTTA  
TGTCTGAAGCAAAATGTTGGACTGAGACTGACCTTACTAAAGGACCTCATGAATTTTGCTCTCAACATAC  
AATGCTAGTTAAACAGGGTGATGATTATGTGTACCTTCCTTACCCAGATCCATCAAGAATCCTAGGGGCC  
GGCTGTTTTGTAGATGATATCGTAAAAACAGATGGTACACTTATGATTGAACGGTTCGTGTCTTTAGCTA  
TAGATGCTTACCCACTTACTAAACATCCTAATCAGGAGTATGCTGATGTCTTTCATTTGTACTTACAATA  
CATAAGAAAGCTACATGATGAGTTAACAGGACACATGTTAGACATGTATTCTGTTATGCTTACTAATGAT  
AACACTTCAAGGTATTGGGAACCTGAGTTTTATGAGGCTATGTACACACCGCATACAGTCTTACAGGCTG  
TTGGGGCTTGTGTTCTTTGCAATTCACAGACTTCATTAAGATGTGGTGCTTGCATACGTAGACCATTCTT  
ATGTTGTAAATGCTGTTACGACCATGTCATATCAACATCACATAAAATTAGTCTTGTCTGTTAATCCGTAT  
GTTTGCAATGCTCCAGGTTGTGATGTCACAGATGTGACTCAACTTTACTTAGGAGGTATGAGCTATTATT  
GTAAATCACATAAACCCATTAGTTTTCCATTGTGTGCTAATGGACAAGTTTTTGGTTTTATATAAAAA  
TACATGTGTTGGTAGCGATAATGTTACTGACTTTAATGCAATTGCAACATGTGACTGGACAAATGCTGGT  
GATTACATTTTAGCTAACACCTGTACTGAAAGACTCAAGCTTTTTGCAGCAGAAACGCTCAAAGCTACTG  
AGGAGACATTTAACTGTCTTATGGTATTGCTACTGTACGTGAAGTGCTGTCTGACAGAGAATTACATCT  
TTCATGGGAAGTTGGTAAACCTAGACCACCCTTAACCGAAATTATGTCTTTACTGGTTATCGTGTAACCT  
AAAAACAGTAAAGTACAAATAGGAGAGTACACCTTTGAAAAAGGTGACTATGGTGATGCTGTTGTTTACC  
GAGGTACAACAACCTTACAAATTAAATGTTGGTGATTATTTTGTGCTGACATCACATACAGTAATGCCATT  
AAGTGCACCTACACTAGTGCCACAAGAGCACTATGTTAGAATTACTGGCTTATACCCAACACTCAATATC  
TCAGATGAGTTTTCTAGCAATGTTGCAAATTATCAAAAGGTTGGTATGCAAAGTATTCTACACTCCAGG  
GACCACCTGGTACTGGTAAGAGTCATTTTGCTATTGGCCTAGCTCTCTACTACCCTTCTGCTCGCATAGT  
GTATACAGCTTGCTCTCATGCCGCTGTTGATGCACTATGTGAGAAGGCATTAAAATATTTGCCTATAGAT  
AAATGTAGTAGAATTATACCTGCACGTGCTCGTGTAGAGTGTTTTGATAAATTCAAAGTGAATTC AACAT  
TAGAACAGTATGTCTTTTGTACTGTAAATGCATTGCCTGAGACGACAGCAGATATAGTTGTCTTTGATGA  
AATTTCAATGGCCACAAATTATGATTTGAGTGTTGTCAATGCCAGATTACGTGCTAAGCACTATGTGTAC  
ATTGGCGACCCTGCTCAATTACCTGCACCACGCACATTGCTAACTAAGGGCACACTAGAACCAGAATATT  
TCAATTCAGTGTGTAGACTTATGAAAACCTATAGGTCCAGACATGTTCCCTCGGAACCTTGTCGGCGTTGTCC  
TGCTGAAATTGTTGACACTGTGAGTGCTTTGGTTTTATGATAATAAGCTTAAAGCACATAAAGACAAATCA  
GCTCAATGCTTTAAATGTTTTATAAGGGTGTTATCACGCATGATGTTTCATCTGCAATTAACAGGCCAC  
AAATAGGCGTGGTAAGAAATTCCCTTACACGTAACCCCTGCTTGGAGAAAAGCTGTCTTTATTTACCTTTA  
TAATTCACAGAATGCTGTAGCCTCAAAGATTTTGGGACTACCAACTCAAAGTGTGATTATCACAGGGC  
TCAGAATATGACTATGTCATATTCACCTCAAACCACTGAAACAGCTCACTCTTGTAATGTAAACAGATTTA  
ATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCATAATGTCTGATAGAGACCTTTATGACAAGTT  
GCAATTTACAAGTCTTGAAATTCACGTAGGAATGTGGCAACTTTACAAGCTGAAAATGTAACAGGACTC  
TTTAAAGATTGTAGTAAGGTAATCACTGGGTACATCCTACACAGGCACCTACACACCTCAGTGTTGACA  
CTAAATTCAAACTGAAGGTTTTATGTGTTGACATACCTGGCATACCTAAGGACATGACCTATAGAAGACT  
CATCTCTATGATGGGTTTTAAATGAATTATCAAGTTAATGGTTACCCTAACATGTTTATCACCCGCGAA  
GAAGCTATAAGACATGTACGTGCATGGATTGGCTTCGATGTCGAGGGGTGTCATGCTACTAGAGAAGCTG  
TTGGTACCAATTTACCTTTACAGCTAGGTTTTTCTACAGGTGTTAACCTAGTTGCTGTACCTACAGGTTA  
TGTTGATACACCTAATAATACAGATTTTTCCAGAGTTAGTGCTAAACCACCGCCTGGAGATCAATTTAAA  
CACCTCATACCCTTATGTACAAAGGACTTCCTTGGAATGTAGTGCGTATAAAGATTGTACAAATGTTAA  
GTGACACACTTAAAAATCTCTCTGACAGAGTCGTATTTGTCTTATGGGCACATGGCTTTGAGTTGACATC  
TATGAAGTATTTTGTGAAAATAGGACCTGAGCGCACCTGTTGTCTATGTGATAGACGTGCCACATGCTTT  
TCCACTGCTTCAGACACTTATGCCTGTTGGCATCATTCTATTGGATTTGATTACGTCTATAATCCGTTTA  
TGATTGATGTTCAACAATGGGGTTTTACAGGTAACCTACAAAGCAACCATGATCTGTATTGTCAAGTCCA  
TGGTAATGCACATGTAGCTAGTTGTGATGCAATCATGACTAGGTGTCTAGCTGTCCACGAGTGCTTTGTT  
AAGCGTGTTGACTGGACTATTGAATATCCTATAATTGGTGATGAAGTGAAGATTAAATGCGGCTTGTAGAA  
AGGTTCAACACATGGTTGTTAAAGCTGCATTATTAGCAGACAAATTTCCAGTTCTTACAGACATTGGTAA  
CCCTAAAGCTATTAAGTGTGTACCTCAAGCTGATGTAGAATGGAAGTTCTATGATGCACAGCCTTGTAGT  
GACAAAGCTTATAAAATAGAAGAATTATTCTATTCTTATGCCACACATTCTGACAAATTCACAGATGGTG  
TATGCCTATTTTGGAAATTGCAATGTGATAGATATCCTGCTAATTCCATTGTTTGTAGATTTGACACTAG  
AGTGCTATCTAACCTTAACCTTGCTGTTGTGATGGTGGCAGTTTGTATGTAAATAAACATGCATTCCAC  
ACACCAGCTTTTGATAAAAGTGCTTTTGTTAATTTAAACAATTACCATTTTTCTATTACTCTGACAGTC  
CATGTGAGTCTCATGGAACAAGTAGTGTCAGATATAGATTATGTACCACTAAAGTCTGCTACGTGTAT  
AACACGTTGCAATTTAGGTGGTGCTGTCTGTAGACATCATGCTAATGAGTACAGATTGTATCTCGATGCT  
TATAACATGATGATCTCAGCTGGCTTTAGCTTGTGGGTTTTACAAACAATTTGATACTTATAACCTCTGGA  
ACACTTTTACAAGACTTCAGAGTTTAGAAAATGTGGCTTTTAAATGTTGTAAATAAGGGACACTTTGATGG  
ACAACAGGGTGAAGTACCAGTTTCTATCATTAATAACACTGTTTACACAAAAGTTGATGGTGTTGATGTA



GAATTGTTTGAAAATAAAACAACATTACCTGTTAATGTAGCATTTGAGCTTTGGGCTAAGCGCAACATTA  
AACCAGTACCAGAGGTGAAAATACTCAATAATTTGGGTGTGGACATTGCTGCTAATACTGTGATCTGGGA  
CTACAAAAGAGATGCTCCAGCACATATATCTACTATTGGTGTGTTTCTATGACTGACATAGCCAAGAAA  
CCAACCTGAAACGATTTGTGCACCACCTCACTGTCTTTTTTGGATGGTAGAGTTGATGGTCAAGTAGACTTAT  
TTAGAAATGCCCCGTAATGGTGTCTTATTACAGAAGGTAGTGTTAAAGGTTTACAACCATCTGTAGGTCC  
CAAACAAGCTAGTCTTAATGGAGTCACATTAATTGGAGAAGCCGTAAAAACACAGTTCAATTATTATAAG  
AAAGTTGATGGTGTGTCCAACAATTACCTGAAACTTACTTTACTCAGAGTAGAAATTTACAAGAATTTA  
AACCCAGGAGTCAAATGGAAATTGATTTCTTAGAATTAGCTATGGATGAATTCATTGAACGGTATAAATT  
AGAAGGCTATGCCTTCGAACATATCGTTTATGGAGATTTTAGTCATAGTCAGTTAGGTGGTTTACATCTA  
CTGATTGGACTAGCTAAACGTTTTAAGGAATCACCTTTTGAATTAGAAGATTTTATTCCCTATGGACAGTA  
CAGTTAAAAACTATTTTCATAACAGATGCGCAACAGGTTTATCTAAGTGTGTGTGTTCTGTTATTGATTT  
ATTACTTGATGATTTTTGTTGAAATAATAAAATCCCAAGATTTATCTGTAGTTTCTAAGGTTGTCAAAGTG  
ACTATTGACTATACAGAAATTTTCAATTTATGCTTTGGTGTAAGATGGCCATGTAGAAACATTTTACCCAA  
AATTACAATCTAGTCAAGCGTGGCAACCGGGTGTGCTATGCCTAATCTTTACAAAATGCAAAGAATGCT  
ATTAGAAAAGTGTGACCTTCAAATTATGGTGATAGTGCAACATTACCTAAAGGCATAATGATGAATGTC  
GCAAAATATACTCAACTGTGTCAATATTTAAACACATTAACATTAGCTGTACCCTATAATATGAGAGTTA  
TACATTTTGGTGCTGGTTCTGATAAAGGAGTTGCACCAGGTACAGCTGTTTTAAGACAGTGGTTGCCTAC  
GGGTACGCTGCTTGTGATTGAGATCTTAATGACTTTGTCTCTGATGCAGATTCAACTTTGATTGGTGAT  
TGTGCAACTGTACATACAGCTAATAAATGGGATCTCATTATTAGTGATATGTACGACCCTAAGACTAAAA  
ATGTTACAAAAGAAAATGACTCTAAAGAGGGTTTTTCACTTACATTTGTGGGTTTATACAACAAAAGCT  
AGCTCTTGGAGGTTCCGTGGCTATAAAGATAACAGAACATTCTTGGAATGCTGATCTTTATAAGCTCATG  
GGACACTTCGCATGGTGGACAGCCTTTGTTACTAATGTGAATGCGTCATCATCTGAAGCATTTTTAATTG  
GATGTAATTATCTTGGCAAACACGCGAACAATAGATGGTTATGTCATGCATGCAAATTACATATTTTG  
GAGGAATACAAATCCAATTCAGTTGTCTTCCCTATTCTTTATTTGACATGAGTAAATTTCCCTTAAATTA  
AGGGGTACTGCTGTTATGTCTTTAAAGAAGGTCAAATCAATGATATGATTTTATCTCTTCTTAGTAAAG  
GTAGACTTATAATTAGAGAAAACAACAGAGTTGTTATTTCTAGTGATGTTCTTGTTAACAACATAACGAA  
CAATGTTTGTGTTTTCTTGTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGAACTCA  
ATTACCCCTGCATACACTAATTCTTTCACACGTGGTGTGTTATTACCCTGACAAAGTTTTTCCAGATCCTCA  
GTTTACATTTCAACTCAGGACTTGTCTTACCTTTCTTTTCCAATGTTACTTGGTTCCATGCTATACATG  
TCTCTGGGACCAATGGTACTAAGAGGTTTGATAAACCTGTCCCTACCATTTAATGATGGTGTGTTTTTGC  
TTCCACTGAGAAGTCTAACATAATAAGAGGCTGGATTTTTTGGTACTACTTTAGATTGGAAGACCCAGTCC  
CTACTTATTGTTAATAACGCTACTAATGTTGTTATTAAAGTCTGTGAATTTCAATTTTGAATGATCCAT  
TTTTGGGTGTTTATTACCACAAAAACAACAAAAGTTGGATGGAAAGTGAGTTCAGAGTTTATTCTAGTGC  
GAATAATTGCACTTTTGAATATGTCTCTCAGCCTTTTCTTATGGACCTTGAAGGAAAACAGGGTAATTTTC  
AAAAATCTTAGGGAATTTGTGTTTAAAGAATATTGATGGTTATTTTAAATATATTCTAAGCACACGCCTA  
TTAATTTAGTGCGTGATCTCCCTCAGGGTTTTTTCGGCTTTAGAACCATTTGGTAGATTTGCCAATAGGTAT  
TAACATCACTAGGTTTCAAACCTTTACTTGCTTTACATAGAAGTTATTTGACTCCTGGTGATTCTTCTTCA  
GGTTGGACAGCTGGTGCTGCAGCTTATTATGTGGGTTATCTTCAACCTAGGACTTTTCTATTAATAATATA  
ATGAAAATGGAACCATTAACAGATGCTGTAGACTGTGCACCTTGACCCTCTCTCAGAAACAAAGTGACGTT  
GAAATCCTTCACTGTAGAAAAAGGAATCTATCAAACCTTCTAAGTTTGAAGTCCAACCAACAGAATCTATT  
GTTAGATTTTCTAATATTACAACTTGTGCCCTTTTGGTGAAGTTTTTAACGCCACCAGATTTGCATCTG  
TTTATGCTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCTATATAATTCCGCATC  
ATTTTCCACTTTTAAAGTGTTATGGAGTGTCTCCTACTAAATTAATGATCTCTGCTTTACTAATGTCTAT  
GCAGATTCATTTGTAATTAGAGGTGATGAAGTCAGACAAATCGCTCCAGGGCAAACCTGGAAAGATTGCTG  
ATTATAATTATAAATTACCAGATGATTTTACAGGCTGCGTTATAGCTTGGAAATCTAACAATCTTGATTCT  
TAAGGTTGGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGTCTAATCTCAAACCTTTTGAAGAG  
GATATTTCAACTGAAATCTATCAGGCCGGTAGCACACCTTGTAATGGTGTGGAAGGTTTTAATTGTTACT  
TTCTTTTACAATCATATGGTTTTCAACCCACTAATGGTGTGTTGTTACCAACCATACAGAGTAGTAGTACT  
TTCTTTTGAACCTTCTACATGCACCAGCAACTGTTTGTGGACCTAAAAAGTCTACTAATTTGGTTAAAAAC  
AAATGTGTCAATTTCAACTTCAATGGTTTAAACAGGCACAGGTGTTCTTACTGAGTCTAACAAAAAGTTTC  
TGCCTTTCCAACAATTTGGCAGAGACATTGCTGACACTACTGATGCTGTCCGTGATCCACAGACACTTGA  
GATTCTTGACATTACACCATGTTCTTTTGGTGGTGTGAGTGTATAACACCAGGAACAAATACTTCTAAC  
CAGGTTGCTGTTCTTTATCAGGATGTTAACTGCACAGAAGTCCCTGTTGCTATTTCATGCAGATCAACTTA  
CTCCTACTTGGCGTGTTTATTCTACAGGTTCTAATGTTTTTCAAACACGTGCAGGCTGTTTAAATAGGGGC  
TGAACATGTCAACAACCTCATATGAGTGTGACATACCCATTGGTGCAGGTATATGCGCTAGTTATCAGACT  
CAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAATCCATCATTGCCTACACTATGTCACCTTG  
GTGCAGAAAATTCAGTTGCTTACTCTAATAACTCTATTGCCATACCCACAAATTTTACTATTAGTGTTAC

CACAGAAATTCTACCAAGTGTCTATGACCAAGACATCAGTAGATTGTACAATGTACATTTGTGGTGATTCA  
ACTGAATGCAGCAATCTTTTGTGCAATATGGCAGTTTTTGTACACAATTAACCGTGCTTTAACTGGAA  
TAGCTGTTGAACAAGACAAAAACCCCAAGAAGTTTTTGCACAAGTCAAACAAATTTACAAAACACCACC  
AATTAAAGATTTTGGTGGTTTTAATTTTTTCACAAATATTACCAGATCCATCAAAACCAAGCAAGAGGTCA  
TTTATTGAAGATCTACTTTTCAACAAAGTGACACTTGCAGATGCTGGCTTCATCAAAACAATATGGTGATT  
GCCTTGGTGATATTGCTGCTAGAGACCTCATTTGTGCACAAAAGTTTAAACGGCCTTACTGTTTTGCCACC  
TTTGCTCACAGATGAAATGATTGCTCAATACACTTCTGCACTGTTAGCGGGTACAATCACTTCTGGTTGG  
ACCTTTGGTGCAGGTGCTGCATTACAAATACCATTTGCTATGCAAATGGCTTATAGGTTTAAATGGTATTG  
GAGTTACACAGAATGTTCTCTATGAGAACCAAAAATTGATTGCCAACCAATTTAATAGTGCTATTGGCAA  
AATTCAAGACTCACTTTCTTCCACAGCAAGTGCACTTGGAAAACCTCAAGATGTGGTCAACCAAAATGCA  
CAAGCTTTAAACACGCTTGTGTTAAACAACCTTAGCTCCAATTTTGGTGCAATTTCAAGTGTTTTAAATGATA  
TCCTTTACCGTCTTGACAAAGTTGAGGCTGAAGTGCAAATTTGATAGGTTGATCACAGGCAGACTTCAAAG  
TTTGAGACATATGTGACTCAACAATTAATTAGAGCTGCAGAAATCAGAGCTTCTGCTAATCTTGCTGCT  
ACTAAAATGTGAGAGTGTGTACTTGGACAATCAAAAAGAGTTGATTTTTGTGGAAAGGGCTATCATCTTA  
TGTCCTTCCCTCAGTCAGCACCTCATGGTGTAGTCTTCTTGCACTGTGACTTATGTCCCTGCACAAGAAAA  
GAACTTCACAACCTGCTCCTGCCATTTGTCTATGATGGAAAAGCACACTTTCCTCGTGAAGGTGTCTTTGTT  
TCAAATGGCACACACTGGTTTGTAAACACAAAGGAATTTTTATGAACCACAAATCATTACTACAGACAACA  
CATTTGTGTCTGGTAACTGTGATGTTGTAATAGGAATTGTCAACAACACAGTTTATGATCCTTTGCAACC  
TGAATTAGACTCATTCAAGGAGGAGTTAGATAAATATTTTAAAGATCATAATCACCAGATGTTGATTTA  
GGTGACATCTCTGGCATTAAATGCTTCAGTTGTAAACATTCAAAAAGAAATTGACCGCCTCAATGAGGTTG  
CCAAGAATTTAAATGAATCTCTCATCGATCTCCAAGAACCTTGAAAGTATGAGCAGTATATAAAATGGCC  
ATGGTACATTTGGCTAGGTTTTATAGCTGGCTTGATTGCCATAGTAATGGTGACAATTATGCTTTGCTGT  
ATGACCAGTTGCTGTAGTTGTCTCAAGGGCTGTTGTTCTTGTGGATCCTGCTGCAAATTTGATGAAGACG  
ACTCTGAGCCAGTGCTCAAAGGAGTCAAATTACATTACACATAAACGAACCTTATGGATTTGTTTATGAGA  
ATCTTCACAATTGGAACCTGTAACCTTTGAAGCAAGGTGAAATCAAGGATGCTACTCCTTCAGATTTTGTTC  
GCGCTACTGCAACGATACCGATACAAGCCTCACTCCCTTTCGGATGGCTTATTGTTGGCGTTGCACTTCT  
TGCTGTTTTTTCAGAGCGCTTCCAAAATCATAACCCTCAAAAAGAGATGGCAACTAGCACTCTCCAAGGGT  
GTTCACTTTGTTTGAACCTTGCTGTTGTTGTTGTTGTTGTAACAGTTTACTCACACCTTTTGCTCGTTGCTGCTG  
GCCTTGAAGCCCCCTTTCTCTATCTTTATGCTTTTAGTCTACTTCTTGCAGAGTATAAACTTTGTAAGAAT  
AATAATGAGGCTTTGGCTTTGCTGGAAATGCCGTTCCAAAACCCATTACTTTATGATGCCAATGATTTTT  
CTTTGCTGGCATACTAATTGTTACGACTATTGTATACCTTACAATAGTGTAACCTTCTTCAATTGTCATTA  
CTTCAGGTGATGGCACAACAAGTCCTATTTCTGAACATGACTACCAGATTGGTGGTTATACTGAAAAATG  
GGAATCTGGAGTAAAGACTGTGTTGTATTACACAGTTACTTCACCTCAGACTATTACCAGCTGTACTCA  
ACTCAATTGAGTACAGACACTGGTGTGTAACATGTTACCTTCTTCATCTACAATAAAATTGTTGATGAGC  
CTGAAGAACATGTCCAAATTCACACAATCGACGGTTCATCCGGAGTTGTTAATCCAGTAATGGAACCAAT  
TTATGATGAACCGACGACGACTACTAGCGTGCCTTTGTAAGCACAAAGCTGATGAGTACGAACCTTATGTAC  
TCATTGTTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCTGTTGAT  
TCTTGCTAGTTACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCTGCAATATTGTTAACGT  
GAGTCTTGTAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGAT  
CTTCTGGTCTAAACGAACATAAATATTATATTAGTTTTTCTGTTTGGAACCTTAAATTTTAGCCATGGCAGA  
TTCCAACGGTACTATTACCGTTGAAGAGCTTAAAAAGCTCCTTGAACAATGGAACCTAGTAATAGGTTTC  
CTATTCCTTACATGGATTTGTCTTCTACAATTTGCCTATGCCAACAGGAATAGGTTTTTGTATATAATTA  
AGTTAATTTTCTCTGGCTGTTATGGCCAGTAACCTTTAGCTTGTTTTGTGCTTGCTGCTGTTTACAGAAT  
AAATTGGATCACCGGTGGAATTGCTATCGCAATGGCTTGCTTGTAGGCTTGATGTGGCTCAGCTACTTC  
ATTGCTTCTTTCAGACTGTTTGCAGCTACGCGTTCATGTGGTCATTCAATCCAGAACTAACATTCTTC  
TCAACGTGCCACTCCATGGCACTATTCTGACCAGACCGCTTCTAGAAAGTGAACCTCGTAATCGGAGCTGT  
GATCCTTTCGTGGACATCTTCGTATTGCTGGACACCATCTAGGACGCTGTGACATCAAGGACCTGCCTAAA  
GAAATCACTGTTGCTACATACGAACGCTTCTTATTACAAATTTGGGAGCTTCGACGCGTGTAGCAGGTG  
ACTCAGGTTTTGCTGCATACAGTCGTACAGGATTGGCAACTATAAATTAACACAGACCATTCCAGTAG  
CAGTGACAATATTGCTTTGCTTGTACAGTAAGTGACAACAGATGTTTCATCTCGTTGACTTTTCAGGTTAC  
TATAGCAGAGATATTACTAATTATTATGAGGACTTTTAAAGTTTCCATTTGGAATCTTGATTACATCATA  
AACCTCATAATTAATAAATTTATCTAAGTCACTAACTGAGAATAAATATTCTCAATTAGATGAAGAGCAAC  
CAATGGAGATTGATTAAACGAACATGAAAATTATTCTTTTCTTGGCACTGATAACACTCGCTACTTGTGA  
GCTTTATCACTACCAAGAGTGTGTTAGAGGTACAACAGTACTTTTAAAAGAACCTTGCTCTTCTGGAACA  
TACGAGGGCAATTCACCATTTTCATCTCTAGCTGATAACAAATTTGCACTGACTTGCTTTAGCACTCAAT  
TTGCTTTTGTCTTGTCTGACGGCGTAAAACACGTCTATCAGTTACGTGCCAGATCAGTTTTCACCTAACT  
GTTTCATCAGACAAGAGGAAGTTCAAGAACCTTTACTCTCCAATTTTTCTTATTGTTGCGGCAATAGTGTTT

ATAACACTTTGCTTCACACTCAAAAGAAAGACAGAATGATTGAACTTTCATTAATTGACTTCTATTTGTG  
CTTTTTAGCCTTTCTGCTATTCCTTGTTTTAATTATGCTTATTATCTTTTGTTTCTCACTTGAAGTGCAA  
GATCATAATGAACTTGTGACGCCTAAACGAACATGAAATTTCTTGTTTTCTTAGGAATCATCACAAGTG  
TAGCTGCATTTACCAAGAATGTAGTTTACAGTCATGTACTCAACATCAACCATATGTAGTTGATGACCC  
GTGTCCTATTCACCTTCTATTCTAAATGGTATATTAGAGTAGGAGCTAGAAAATCAGCACCTTTAATTGAA  
TTGTGCGTGGATGAGGCTGGTTCTAAATCACCCATTTCAGTACATCGATATCGGTAATTATACAGTTTCCT  
GTTTACCTTTTACAATTAATTGCCAGGAACCTAAATTGGGTAGTCTTGTAGTGCGTTGTTTCGTTCTATGA  
AGACTTTTTAGAGTATCATGACGTTTCGTGTTGTTTTAGATTTTCATCTAAACGAACAAACTAAAATGTCTG  
ATAATGGACCCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCCTCAGATTCAACTGGCAG  
TAACCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACT  
GCGTCTTGGTTCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCTCGAGGACAAGGCGTTC  
CAATTAACACCAATAGCAGTCCAGATGACCAAAATTGGCTACTACCGAAGAGCTACCAGACGAATTTCGTGG  
TGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAACTGGGCCAGAAGCT  
GGACTTCCCTATGGTGCTAACAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAATACACCAA  
AAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACCTTCTCAAGGAACAAC  
ATTGCCAAAAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTTCCTCATCACGT  
AGTCGCAACAGTTCAAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCA  
ATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGACAGATTGAACCAGCTTGAGAGCAAAATGTCTGG  
TAAAGGCCAACAACAACAAGGCCAAACTGTCTACTAAGAAATCTGCTGCTGAGGCTTCTAAGAAGCCTCGG  
CAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAACCC  
AAGGAAATTTTGGGGACCAGGAATAATCAGACAAGGAAGTATTACAAACATTGGCCGCAAATTGCACA  
ATTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATGTGCGGCATTGGCATGGAAGTCACACCTTCGGGAACG  
TGGTTGACCTACACAGGTGCCATCAAATTGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGC  
TGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGC  
TGATGAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAAGTGTGACTCTTCTTCCTGCTGCAGAT  
TTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAAACTCATG  
CAGACCACACAAGGCAGATGGGCTATATAAACGTTTTTCGCTTTTCCGTTTACGATATATAGTCTACTCTT  
GTGCAGAATGAATTCTCGTAACTACATAGCACAAAGTAGATGTAGTTAACTTTAATCTCACATAGCAATCT  
TTAATCAGTGTGTAACATTAGGGAGGACTTGAAAGAGCCACCACATTTTCACCGAGGCCACGCGGAGTAC  
GATCGAGTGACAGTGAACAATGCTAGGGAGAGCTGCCTATATGGAAGAGCCCTAATGTGTAAAATTAAT  
TTTAGTAGTGCTATCCCCATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAA  
AAAAAAAAAAAAA

## **Appendix B: reads to be used in testing**

```
>read_01
AGGGTTCAGGAAAGAGTTTGGGAAGGATGCTGAACATCTTGAATAGGAGAC
>read_02
AAGACACAATCACGTACAAGTGTGCCCTTCTCAGGCAGAATGAGCCAGAA
>read_03
AAGGGGGCTGGAAACATGTCCAGAGAATTGAAACTTGGATCTTGAGACAT
>read_04
AGAGGCTAAGCTAACCAACACAACAACAGAATCTCGCTGCCCCAACACAAG
>read_05
AATGACACAGGGAAACATGGCAAGGAAATCAAAATAACACCACAGAGTTC
>read_06
ATGTTACGAAAGAAAATGACTCTAAAGAGGGTTTTTTTCACTTACATTTGT
>read_07
AGCTCTTGGAGGTTCCGTGGCTATAAAGATAACGGAACATTCTTGGAATG
>read_08
GGACACTTCGCATGGTGGACAGCCTTTGTTACTAATGTGAATGCGTCATC
>read_09
GATGTAATTATCTTGGCTAACCACGCGAACAAATTGATGGTTATGTTATG
>read_10
GAGGAATACCAATCCAATTCACTTGTCTCCCTATTCTTTCTTTGACATGA
>read_11
AGGGGTACTGCTGTTATGTCTTTAGAAGGTCAAATCAATGATATGATCCC
>read_12
GTAGACTTATAATTAGAGAAAACAACAGAGTTGTTATTTCTAGTGATGTT
>read_13
CAATGTTTGTTTTTCTTGCCCTTTTATTGCCACTAGTCTCTAGTCAGTGTG
>read_14
ATTACCCCCTGCATACACTAGGGATTCTTTCACACGTGGTGTTTATTACC
>read_15
GTTTTACATTCAACTCAGGACTTGTCTTACCTTTCTTTTCCAATGTTAC
```