T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

COMPUTER ENGINEERING DEPARTMENT

CSE 4065

Computational Genomics

Programming Assignment # 2

Group Members

Abdulkerim Talha Timur - 150119657

Supervised by

Assoc. Prof. Betül Boz

Introduction

This project focuses on implementing a dynamic pairwise sequence alignment algorithm, specifically the Needleman-Wunsch algorithm. Sequence alignment is crucial in bioinformatics for comparing DNA sequences, which helps in understanding genetic relationships, evolutionary histories, and functional genomics. The goal of this project is to align DNA sequences optimally using predefined scoring systems for matches, mismatches, and gaps.

Methods

Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm is a globally recognized method for sequence alignment. It uses a dynamic programming approach to ensure that the optimal alignment between two sequences is found by considering every possible alignment and scoring them based on a predefined scoring system.

Scoring System

- Match Score: A positive score (+7) is given when two nucleotides from the sequences being aligned are the same.
- Mismatch Score: A negative score (-5) is applied when two nucleotides do not match.
- Affine Gap Penalty: This includes a gap opening penalty (-3) and a gap extension penalty (-1), which penalize the alignment for starting a gap and extending it, respectively.

Project File Structure and Modules

- **main.py:** This script manages the flow of the program. It coordinates the reading of sequences from input files, invokes the alignment processes, and outputs the results.
- **alignment.py:** Contains the core functions of the Needleman-Wunsch algorithm and the traceback function to determine the optimal alignment path.
- **file_manager.py:** Handles reading of DNA sequences from the given input files located in a specified directory.
- **matrix.py:** Responsible for initializing the matrices used in the alignment algorithm, setting up the necessary structure for dynamic programming.

Results

File: test1.seq

Alignment Score: 406

Sequence 1:

```
---C-GAGAC-C-GACG--AAGAGGT--T-T--GG----C-C----CCAACCAGG---TTC-C-C-T----GA------TCACG-T--A-A-CTTACCG-----G
C-C--AA--AAGGACT----GGCC-TTACT------AA-G-GC--CTTTGTC-T-A-C-T-GC-GGGTCCGGGGG-C-C-G-T---TGG-T-TTCGGC
AGAACAC-----TT-C
```

Sequence 2:

```
CCT-GGAC-CG-AG-C-TTAA-AT-TGCTA-GC--AATACA-GATGCCG--C---TTC--CT-T-G-GGGAG-GGTGTGT----A-GGAT-G--TA
G-GTTAACG-A-AT--GCAAGT--TCCGGGGT-A-T-C-GCAGAG--T-CG-TGC----T-A-C-G-G-CG-T-GG--C-----ACT-T-A-GGGT--CTC
-TCGG--GAAAAAGAGTA--G-
```

File: test2.seq

Alignment Score: 725

Sequence 1:

```
-GTGTGGTTGCT-TGCATCACTCCGTGTACA-TGTG-----AC-----AAC---CGAAC--GA-GT-T-GA--T----CC-A--GCTTTGTTAAGT
CAG-CTT--CGAATGCGGTAGCTCTCAAATATGATATGAC-TCTTG-GG-GTAGATG-CTGGGGACCTATTGCG-CCC-AAAG-
CGATAT-TC-G-------GGGC-ACCGGTTTAG-GGTACCCTATCAA----GGC-G----ATACTTCGAT-G--C-AG-TG-T-GATG-C-CGGA
GGTGTC-GG-T-C-AG-CATGT-G-AG-A-G-CT
```

Sequence 2:

```
GGTA-GGTT-C-GTGAAGCACTCC-TGGACTCTGA-CCACAACATAAT--CAAGCGCA-GAG-T-TG-A-ATG-ACCA--AAGCGC--
--TCAAG-C--T-TTCACGAA--C----G-TCTC-------ATATC-CG-CG-GT--CGTACAA-AC----G--C----GC-T-CTTAAA-ACAATC-G--T-
CTATCCA----T-CCGGG---GA--TACCA-AT---GGGT-G-A-CATTA-AC-T-G--GGCA-G-GC--C-GGC--ACGCGGACGCG--A--C-A-G
A-T--TG-A-AA-T-GGA-T
```

File: test3.seq

Alignment Score: 2418
Sequence 1:

```
-------CGGGGA-AAG-AC---GG---A-A-T-G-CA-T-C-GACCATCGGAC--A-AT--G-CCTCAC--TG--GAAGC---GCT--GCTG-ATTTTT-
---G-C-GC-AACGA-GC-CT-C-G-GA---CCTCC-CG-CT-CAA-ACTT---AC-GA-A-AATGA---CT---CCAC-C------A-G--CAC-T-GA-A-
CCAACTGG--C-CTC-CA-GT-G-A-A-G-A--------T--AG-TA-T-C-T-AC-A-AA--TC--G----TT--TG----CC---GG-GAGA-AA----C----AT-T
TGTATGGTAGTCAGCG--T-T-C---TGCACGTCACGTAATCGTTCACTAGTTGTGGGGTAT-ACCAGGTCGTAATGAGAT-G-T
TA-G-T-A-TGAATCGTTTATAA---CG-C-T-T-GTCAT--AGGA-G-T-----TCC--G-A-ATAATCG--TCA-CT--AG-G-T-C-A-A-TG--GCC
CCCTA-C---TTG-TA-ATAA-CTACGT---CTGATTGA-GAAA------CAGATCGTTAG-T-CATCGG-T-TAATAGT-CCCGG-A-AGA-T
```

```
--AACGGGT-TCTTGCGTTTTTGCGAATACTACTA-TCT---C------AT---GG-C-G-A-TG--GA-G-CGTTT---G-GT-CCCA-TCCAGC
C---G-CGCGAGTATCA-CT----TGTTCGCCTG-CA-----CC-TG--T-C-CG-A----C-----CTTTCGATGG-----G-G-AGCTCCTTTCA-T-T-
GG-T---TGT-AG-GTACTAAGGGTGAGCAATG--T-CA-CGT---G-ACCCA-AGGAGCCGTGTGTA-ATTTCCACTTGCTCAGAAA
-AG---C-CT--C----G-AC-AAT---C-T-G-GGA----CC-G-A-CACC--TGA-GT-G--AT-GGCTT--AC-A--G-ACC-AGGGG-A---G-GG-GGG-
C-A-G-G-TT----C-CC-T-G---GCCACAGAATGGCA-CG-CC-C-TG-AGGAGGC-ACCG-G-CCCAA---C-G-T-CGCTGG
```

Sequence 2:

```
AGTGAGA-----AG-A-C-CGGA--TATAT-G-C-A-AC-G-A-ACAATCGCA-AAATA-GC-T-C-CA-TGT-ACGC--CTGT-C-CCG-T-AAT
A---CCGC-A-TGCG-A-G-C-CTC-G-A-AG-CCCCCG--ACGTC-A-AAC-CTCAACA-AG-TAT-A-G-GGT-TCCA-C-CG-TATGAG-C
-CACA-G-A-AT-T-C-AGT--TCCTC-CGC-TG-A-TAG-T-C-CCGGTTCCTCCA-AT-C-A-G-CA-GAT--TCT-GA-CTTA--GGT-CACT--
GGA--AGAAAC--CTGGCGTAT-TG-TG-AT-G-A---AA--TTTC-G-TGGTGGACG---C-TAC-CG-T--CTAG--GTA--G-A-G-CCA---C-
T-A-GTG--C-ATTGT-A-CACTG--TC-TTTTTC-CGGC-T-ATA-A-TC--GGA-G-T-TTCAGGCT--GA-C-T-TG--C-CA-C-CC-AG-GT-
C-A-AATAGT-ATGTCA-C--G-GTG-T-AT-GATC-TCT-C--GGGCTGATAG-C-AAGGGGCGGC-G-TCGG--GC-AC-TCC-TTG-AA
TTG-A-----C-CAT-TTGG--CGTG-C-CTT--G---TTGCC--TCCT--TCCTC-AAGCAGGGTTA-GGC--T-T-CAGT-AAG-TGT-G---GGG
GTG-GCT--G-CC-G--AAAGACG-G-G---C-TC-GGGGTGTT-GCCC-A-ACGTCT--G-GTC-G-A--C-TTCCCATAATC--T-G--GGCA
CAAGA-C-G-TA---T--G-G-A--C-CCGTG-C-GC-T--TAATGGT--GC-ATTTC-AC-GCG-AGA-AACG--GA-G-GCAG-GTGC-CA--TC
CGCG-GCG-A-AAAGA-TCA-TCAGACATGA-CA-T-A-ACG-A-GGA--ACCCG--A-T-CCG--GA-G-TG-A-ATA-CGTC--ACA-T-GC
-CA--G-GGGGC-GGTGT-GC---A-G-T-T-C-TCGACCG--G-A-AACGCC-C---A---C-CC-A-CG-A--C-GTAC-CG--C-C-A---AAGCG-
T-C-GC-CTGG
```

File: test4.seq

Alignment Score: 344

Sequence 1:

```
-----AGGCCGAA-AACGTCGCGAATTGACCCTGGCGACGCCGCCGAACGGGACCTCCG-TTA---GTG--T--GGGA-G-G-T--CA
TCAATCT-C--GTT-C--GCTAG-CGGCTGACACCAATCACTATAAGTC-TGTCATGAC
```

Sequence 2:

```
CTTCAAGT-CAAAT-A--T-----A--GATCCTGGC--CGCT-CC--ACGGG--CTT--A--AGTCGT-TCTCCGA-AGGT-A-CG-ATC--TG-G-
TTG--G-ATGCTT-CCGTCT-A-AACAAG-A--AG-A-TAA--TC--G--
```

File: test5.seq

Alignment Score: 548

Sequence 1:

```
----CGG-GTAGTTAACCC-T--AC-AG--C-AT-AGA-G-T-CGC-G-AG-ATA---AAGTGCAG-GAGTCTTTCGCGGCAGATTCGT-A---
CCT-C-A--ACCACGT--GCTACT---T-T-CTGG-CATC-ACGA-ATCTGCCGC-AT-A-G-GTCC-GTGA-GTCCATATGA
```

Sequence 2:

```
AGGA---AGTAGTTAGCCTA-ACA-G-GCA-TA-GAG-T-C-GCG-A-CA-TAT-GTGAA--G-A-T--GTCATT--CGGT--ATTCA-AACCT
-C-A-T-GCATCAT-TGC-CT--TGAGTC-GC---TC--CT--G-GA---GCA--TA-G-T-C--CCTGAG-TG-CCATATGA
```

Discussion

Evaluation of Results

The alignments generated by the program are critical for understanding the similarities and differences between the DNA sequences. By analyzing these alignments, researchers can infer evolutionary relationships or predict functional similarities.

Challenges and Learnings

One of the primary challenges faced during the project was ensuring the accuracy of the alignment under various conditions, such as sequences of different lengths or compositions. The project provided valuable insights into the complexity of biological data analysis and enhanced understanding of dynamic programming in algorithm design.