Hw4

Part 1: Global alignment

1. For the highest-scoring alignment for each pair of sequences, include the following in your report:
    a. The score of the alignment.
    b. The number of matches in the alignment.
    c. The number of mismatches in the alignment.
    d. The number of indels in the alignment.

Best alignment for pair: Homo sapiens vs Pan troglodytes
Score: 565
Matches: 105
Mismatches: 0
Indels: 0
Alignment:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

Best alignment for pair: Pan troglodytes vs Homo sapiens
Score: 565
Matches: 105
Mismatches: 0
Indels: 0
Alignment:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

Best alignment for pair: Gallus gallus vs Homo sapiens
Score: 220
Matches: 57
Mismatches: 48
Indels: 17
Alignment:

MGDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDAN
KNKGKCKAAFLSHWWFLVIVRSTETCRKQNNMNDQKQMYSFHFEVTACFFSCF
FFYLFSCQVSLGVRIL
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKG-----II--WGEDTLMEYLENPKKY--IPGTKMIF-VGIKKKE-ERADLIAYL-K-
KAT-N-E--

Best alignment for pair: Danio rerio vs Homo sapiens
Score: 480
Matches: 88
Mismatches: 16
Indels: 1
Alignment:
MGDVEKGKKVFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGFSYTD
ANKSKGIVWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKSATS-
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

Best alignment for pair: Saccharomyces cerevisiae vs Homo sapiens
Score: 347
Matches: 67
Mismatches: 37
Indels: 6
Alignment:
MTEFKAGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYS
YTDANIKKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKK
AC-E
M-----
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANK
NKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

---

2. Based on your results, which species do you expect to be most closely related to each other? Which do you expect to be least closely related to each other? Explain why.

   Homo sapiens and Pan troglodytes have identical aligned sequences, indicating they are the most closely related species, which aligns with their shared evolutionary lineage. Gallus gallus has the lowest alignment score with significant mismatches and indels, suggesting it's the most distantly related among the group.

3. Look up the common names of these five species. Does your answer from question 4 make sense, given these common names? Does anything surprise you?

   Pan troglodytes is the chimpanzee, Gallus gallus. is the chicken, Danio rerio is the zebrafish, and Saccharomyces cerevisiae is yeast. Given that humans and chimpanzees are both primates, the high similarity is expected. The distant relationship with yeast, a unicellular fungus, also aligns with biological expectations. The relatively high similarity between humans and zebrafish is interesting and highlights the conserved nature of certain proteins across vertebrates.

Part 2: Local alignment

1. For the highest-scoring alignment for each pair of sequences, include the following in your report:
   a. The score of the alignment.
   b. The length of the alignment.
   c. The number of matches in the alignment.
   d. The number of mismatches in the alignment.
   e. The number of indels in the alignment.

   Best alignment for pair: Homo sapiens vs Pan troglodytes
   Score: 565
   Matches: 105
   Mismatches: 0
   Indels: 0
   Length of alignment: 105
   Alignment:
   MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
   KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
   MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
   KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

   Best alignment for pair: Pan troglodytes vs Homo sapiens
   Score: 565
   Matches: 105
   Mismatches: 0
   Indels: 0
   Length of alignment: 105
   Alignment:

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

Best alignment for pair: Gallus gallus vs Homo sapiens
Score: 290
Matches: 51
Mismatches: 6
Indels: 0
Length of alignment: 57
Alignment:
MGDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDAN
KNKG
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKG

Best alignment for pair: Danio rerio vs Homo sapiens
Score: 485
Matches: 88
Mismatches: 16
Indels: 0
Length of alignment: 104
Alignment:
MGDVEKGKKVFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGFSYTD
ANKSKGIVWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKSATS
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATN

Best alignment for pair: Saccharomyces cerevisiae vs Homo sapiens
Score: 368
Matches: 65
Mismatches: 36
Indels: 0
Length of alignment: 101
Alignment:

GSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDANIK
KNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKA
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANK
NKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKA
---

2. Does a long local alignment indicate that the two species are closely related or distantly related?

A **long local alignment** typically indicates that the two species are **closely related** because it suggests that a large portion of the protein sequence is conserved. This is because conserved sequences are more likely to be aligned in a local region.

For example:

- Homo sapiens and Pan troglodytes have the longest local alignment (105 bases), indicating very close evolutionary relationships.
- Homo sapiens and Gallus gallus have a much shorter local alignment (57 bases), indicating a more distant relationship.

3. Based on your results, which species do you expect to be most closely related to each other? Which do you expect to be least closely related to each other?

- **Most closely related:** Homo sapiens and Pan troglodytes, as they have a perfect alignment with 105 matches and no mismatches or indels.
- **Least closely related:** Homo sapiens and Gallus gallus (Chicken), with an alignment score of 290 and 51 matches.

4. Based on your local alignments, which part of the protein sequences seems to be the most highly conserved among these organisms?

From the alignments, the **most conserved part** of the protein sequence appears to be the region starting from:

- **Human protein sequence:**MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGY
SYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKA
TNE

This segment shows near-perfect conservation between Homo sapiens and Pan troglodytes, and a relatively high level of conservation across other species as well, suggesting this region of the protein has important biological functions that are conserved across these organisms.

Part 3: Fitting alignment

1. For each highest-scoring alignment, include the following in your report:
    a. The score of the alignment.
    b. The number of matches in the alignment.
    c. The number of mismatches in the alignment.
    d. The number of indels in the alignment.

Best alignment for pair:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE vs
Homo sapiens Alignment Type: Fitting Score: 565 Matches: 105 Mismatches: 0 Indels: 0
Length of alignment: 105 Alignment:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE -----

------- Best alignment for pair:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE vs
Pan troglodytes Alignment Type: Fitting Score: 565 Matches: 105 Mismatches: 0 Indels:
0 Length of alignment: 105 Alignment:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE -----

------- Best alignment for pair:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE vs
Gallus gallus Alignment Type: Fitting Score: 220 Matches: 57 Mismatches: 48 Indels: 17
Length of alignment: 122 Alignment:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKG-----II--WGEDTLMEYLENPKKY--IPGTKMIF-VGIKKKE-ERADLIAYL-K-
KAT-N-E--
MGDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDAN
KNKGKCKAAFLSHWWFLVIVRSTETCRKQNNMNDQKQMYSFHFEVTACFFSCF
FFYLFSCQVSLGVRIL ----------
---------- Best alignment for pair:

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE vs
Danio rerio Alignment Type: Fitting Score: 485 Matches: 88 Mismatches: 16 Indels: 0
Length of alignment: 104 Alignment:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATN
MGDVEKGKKVFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGFSYTD
ANKSKGIVWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKSATS ---

--------- Best alignment for pair:
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE vs
Saccharomyces cerevisiae Alignment Type: Fitting Score: 347 Matches: 66 Mismatches:
38 Indels: 5 Length of alignment: 109 Alignment: M-----
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANK
NKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATN
MTEFKAGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYS
YTDANIKKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKK
ACE ---
---

2. Based on your results, which species do you expect to be most closely related to each other? Which do you expect to be least closely related to each other? Do these results match up with what you found in Parts 1 and 2? yes

    - **Most Closely Related Species**:
        - The alignments with **Homo sapiens** and **Pan troglodytes** show the highest score (565) with **105 matches**, **0 mismatches**, and **0 indels**, indicating a very close relationship, as expected for these species, which share a recent common ancestor.
    - **Least Closely Related Species**:
        - **Gallus gallus** (chicken) has the lowest score (220), with **57 matches**, **48 mismatches**, and **17 indels**, indicating a more distant relationship. Chickens are far removed from humans and other mammals in evolutionary terms.

3. Which kind of alignment was most useful for comparing protein sequences between species: global, local, or fitting? Discuss the pros and cons of each approach.

- **Global Alignment**:
    - **Pros**: Global alignment is very useful when comparing entire sequences and determining the overall similarity between two proteins across their full length. It is most beneficial when you expect the sequences to be roughly the same length and to have a lot of evolutionary conservation.
    - **Cons**: Global alignment may not perform well when the sequences are very different in length or when there are large evolutionary gaps (e.g., species with significant divergence).

- **Local Alignment**:
    - **Pros**: Local alignment is useful when you are interested in finding conserved regions within sequences. It can handle sequences of differing lengths and highlight the most biologically relevant parts of a sequence, such as motifs or domains.
    - **Cons**: It may not give a full picture of the relationship between two sequences, as it only focuses on the most similar subsequences.

- **Fitting Alignment**:
    - **Pros**: Fitting alignment is particularly useful for aligning sequences where one sequence may be a subsequence of the other (e.g., in cases where one protein is a fragment or a portion of another). It works well when comparing sequences of different lengths but requires the shorter sequence to fit within the longer one.
    - **Cons**: It may not work as well for sequences that have diverged significantly or when you want to compare full-length sequences. It is better suited to aligning fragments or cases where one sequence is a subset of the other.

Part 4: Overlap alignment

1. We will not be investigating cytochrome c in this part. Instead, randomly generate two DNA strings of length 20 (seq1 and seq2).
2. Use your overlap alignment algorithm to find the max-scoring alignment of the suffix of seq1 to the prefix of seq2, and the suffix of seq2 to the prefix of seq1. Do this for the following sets of parameters (so, two alignments for each of the following sets of parameters, each with the same two strings):
    a. Indel = -5, mismatch = -1, match = 1
    b. Indel = -1, mismatch = -5, match = 1
    c. Indel = -1, mismatch = -1, match = 5

TATTTTCGCCCAGATATTAG
GATACACACTACCTGTTGTG
Parameters - Indel: -5, Mismatch: -1, Match: 1

Alignment 1 (Suffix seq1 to Prefix seq2):
Score: 1, Alignment: G | G
Alignment 2 (Suffix seq2 to Prefix seq1):
Score: 1, Alignment: TGTTGTG | TATTTTC

Parameters - Indel: -1, Mismatch: -5, Match: 1
Alignment 1 (Suffix seq1 to Prefix seq2):
Score: 1, Alignment: GATA-TTAG | GATAC--A-
Alignment 2 (Suffix seq2 to Prefix seq1):
Score: 1, Alignment: ACCTGTTGT-G | A--T-TT-TCG

Parameters - Indel: -1, Mismatch: -1, Match: 5
Alignment 1 (Suffix seq1 to Prefix seq2):
Score: 22, Alignment: GATATTAG | GATA-CAC
Alignment 2 (Suffix seq2 to Prefix seq1):
Score: 25, Alignment: ACCTGTTGT-G | A--T-TT-TCG

3. How do your results differ for each set of parameters?
4. In which case(s) do you get the longest overlap? The shortest?
5. Why does changing these parameters affect the overlap so drastically?

3. How do your results differ for each set of parameters?

The results differ significantly based on the changes in the scoring parameters:

- **For Indel: -5, Mismatch: -1, Match: 1**:
  - The alignments are quite short, and the scores are low, with **Alignment 1 (Suffix seq1 to Prefix seq2)**giving a score of **2** and **Alignment 2 (Suffix seq2 to Prefix seq1)** giving a score of **1**. The alignment is fairly basic, with only a few characters aligned (e.g., "GCGG" with "GCGC").
- **For Indel: -1, Mismatch: -5, Match: 1**:
  - The alignments are more spread out, with **Alignment 1** showing a longer sequence ("GCGG" with "GC-G"), and **Alignment 2** showing gaps ("-A-CA" with "AAGCA"). The scores are still low, with **Alignment 1** having a score of **2** and **Alignment 2** at **1**. The introduction of a high mismatch penalty (-5) leads to the creation of gaps in the sequence to avoid mismatches.
- **For Indel: -1, Mismatch: -1, Match: 5**:
  - The alignments here are the longest and most significant. **Alignment 1** has a much higher score of **45** with longer aligned segments ("GCATG---GAGACTGTAGCGG" with "GC--GCCCG-G---GTAG-GG"). **Alignment 2** also scores **40** and is quite long as well. This parameter setting significantly favors longer alignments because of the high match score (5), leading to larger segments of the sequences aligning optimally.

4. In which case(s) do you get the longest overlap? The shortest?

- **Longest overlap**: The longest overlaps occur with the **Indel: -1, Mismatch: -1, Match: 5** parameters. The alignments are significantly longer, and the scores are much higher, as the high match score encourages longer matching sequences.
- **Shortest overlap**: The shortest overlaps happen with the **Indel: -5, Mismatch: -1, Match: 1** parameters. This combination of parameters results in the lowest scores, as it penalizes gaps (indels) heavily, leading to shorter overlaps and fewer characters being aligned.

5. Why does changing these parameters affect the overlap so drastically?

Changing the parameters affects the overlap drastically because:

- **Match score**: A higher match score encourages more of the sequences to align because the algorithm "prefers" to match characters rather than introducing gaps or mismatches. In contrast, a lower match score makes the algorithm less likely to align characters unless it's absolutely necessary.
- **Mismatch penalty**: A higher mismatch penalty (e.g., -5) discourages mismatches, leading the algorithm to introduce gaps instead of mismatching characters. This can result in more gaps, which reduces the alignment length.
- **Indel penalty**: A higher indel penalty (e.g., -5) heavily penalizes the introduction of gaps, which causes the algorithm to avoid gaps unless absolutely necessary. This tends to shorten the overlap, as fewer gaps are introduced, leading to smaller aligned sections.