Jaehwi Kim

jkim589

CS466

December 16, 2018

Gene/Protein Sequence Alignment Report

* **Brief Motivation**

In bioinformatics, sequence alignment is used to arrange sequences of DNA, RNA, or protein to identify regions of similarity. This alignment may indicate the functional, structural, evolutionary significance of the sequence. In this project, I will analyze gene and protein sequences of three different organisms from Primates family using global and local pair-wise sequence alignments. Within the Primates family, I will compare sequence alignments of Human and Chimpanzee with the alignments of Human and Baboon. I expect that Chimpanzees will have more similarities in genes and proteins, and that the sequence alignments of Human and Chimpanzee will have high score and consist of more matches on the sequence alignments. I derive datasets from PFAM, KEGG GENES database and NCBI database to query specific gene and protein families that I wish to compare.

* **Method**

I have used global and local pair-wise sequence alignment to compare two different gene and protein sequences, and particularly used Needleman-Wunsch algorithm for global alignment problem, and Smith-Waterman algorithm for local alignment problem.

* + **Needleman-Wunsch**

Needleman-Wunsch Algorithm was developed by Saul B. Needleman and Christian D. Wunsch in 1970. This algorithm uses dynamic programing to find the optimal global alignment.

Given assigned scores and gap penalty, this algorithm will start from first sequence string in the top and move through the cells row by row, calculating the score for each cell. There are three different paths for calculating scores: diagonal path for representing match/mismatch, and vertical / horizontal path for representing the indel pairing. The score in the cell on the bottom right represent the result alignment score for the optimal alignment.

* + **Smith-Waterman**

Smith-Waterman Algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981. This algorithm also uses dynamic programming to find segments of all possible length that optimize the similarity measure.

Given assigned scores and gap penalty, this algorithm will also start from first sequence string in the top and move through the cells row by row, calculating the score for each cell. The subtle difference in Smith-Waterman Algorithm is that if none of the scores are positive, the element will get a score of 0. (if there is no similarity, the element will be 0). The optimal alignment score will be the highest element present in the dynamic programming table.

* **Validation / Benchmarking Results**

Both Needleman-Wunsch and Smith-Waterman algorithms were run on several different gene/protein sequences I have queried from databases. Each of these sequences varied by length, and chromosome parent so that the benchmarking would cover extensively. The gene/protein sequences that were tested were 12S Ribosomal RNA, COX1, CYTB, ND4L, RLIM, tRNA-Phe, and SYS1. For each of Primates families: Homo Sapiens (Human), Papio Anubis (Baboon), Pan Troglodytes (Chimpanzee), I have executed the program to get pair-wise alignment of Human and Baboon, and Human and Chimpanzee. With these pair-wise alignments, I could compare the optimized scores for each gene/protein sequences, and analyze the genetic similarities / differences between human and two species of primates.

**Needleman-Wunsch Algorithm (Global Alignment)**

|  |  |  |
| --- | --- | --- |
| Family /  Method  Genome /  Protein | Human-Baboon  Needleman-Wunsch | Human-Chimpanzee Needleman-Wunsch |
| 12S Ribosomal RNA | 560 | 847 |
| COX1 | 956 | 1299 |
| CYTB | 703 | 901 |
| ND4L | 175 | 256 |
| RLIM | 397 | 512 |
| tRNA-Phe | 50 | 63 |
| SYS1 | 159 | 157 |

**Smith-Waterman Algorithm / Local Alignment**

|  |  |  |
| --- | --- | --- |
| Family /  Method  Genome /  Protein | Human-Baboon  Smith-Waterman | Human-Chimpanzee  Smith-Waterman |
| 12S Ribosomal RNA | 848 | 848 |
| COX1 | 980 | 1300 |
| CYTB | 703 | 902 |
| ND4L | 175 | 257 |
| RLIM | 507 | 512 |
| tRNA-Phe | 50 | 63 |
| SYS1 | 159 | 157 |

* **Conclusion**

Between the sequence alignments from Smith-Waterman (local alignment) and Needleman-Wunsch (global alignment) Algorithms, the Smith-Waterman Algorithm generally had higher optimized score than the Needleman-Wunsch Algorithm. From analyzing the optimized scores between pair-wise alignment of Human & Baboon and Human & Chimpanzee, the optimized scores were almost consistently higher for Human & Chimpanzee alignment than Human & Baboon alignment. As I have hypothesized that Human & Chimpanzee alignment would have higher score since they have more similar resemblance than the Human & Baboon, the genetical relationship between Human and Chimpanzees also proved that they were more alike than Baboons.

Although I have successfully proved that Human and Chimpanzees are alike in genetic / protein sequences, I was dissatisfied with how my implementation worked to demonstrate the sequence alignments. First, when I have tested my program on several made-up genetic / protein sequences to verify if my implementation worked correctly, I disregarded the situation where genetic / protein sequences were really long. As a result, although my program successfully created neat dynamic programming table and sequence alignment results, it was difficult to analyze the result when real-data long sequences were tested. For future implementation and improvements, I intend to modify the format of output / results so that the users could view the dynamic programming table and sequence alignment with ease without the need of modifying the result file.

There were few additional implementations that could complement my program as well. Adding a function to calculate the runtime of both algorithms would be an interesting addition to my program and allow me to compare the time efficiency of both algorithms and find the most time efficient alignment algorithm. Lastly, using neighbor joining to compute a phylogeny and compare to tree of life would also complement my program as well. Using neighbor joining, I could actually derive my sequences to the tree of life and compute the distance between Human and two other species of Primates and compare these distances to see analyze the likeness of these Primates as well. I plan to keep modifying my program and research more genetic / protein sequences so that they will execute more functions I want, and test on many other genetic questions I have in future.