CS-415: Computational Biology: Project 4

Taylor Martin
University of Idaho
Moscow, Idaho, USA
mart8517@vandals.uidaho.edu

1 ABSTRACT

This project utilized genome alignment techniques from the field of computational biology, specifically as taught in CS-415. The aim was to align a set of provided genome sequences and evaluate the resulting alignments based on a scoring system. To achieve this goal, a substitution matrix generated in a previous project was utilized, along with the Blossum50 matrix. The methodology employed in this investigation yielded satisfactory outcomes, and the confidence level in the results generated is moderate.

2 METHODS

Below are the different methods that I utilized to align the provided genome sequences and generate the alignment scoring results.

2.1 Data Preparation

To prepare the data for genome sequence alignment and scoring, a custom Python module called DataLoader was created. The load_data() function within this module was used to import the eight provided genome sequences, the substitution matrix generated in Project 3, and the Blossum50 matrix. Once loaded, the data was ready for use in subsequent analysis.

2.1.1 Blossum50 Matrix. Below is a figure showing the .json format of the Blossum50 Matrix utilized in this project.

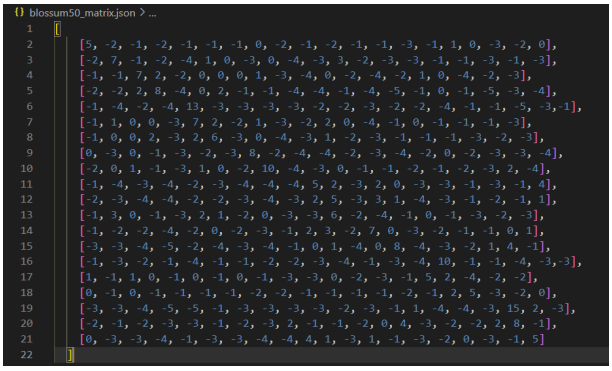

Figure 1: Blossum50 Matrix

2.1.2 Substitution Matrix. Below is a figure showing the .json format of the Substitution Matrix I generated in Project 3 for use in this project.

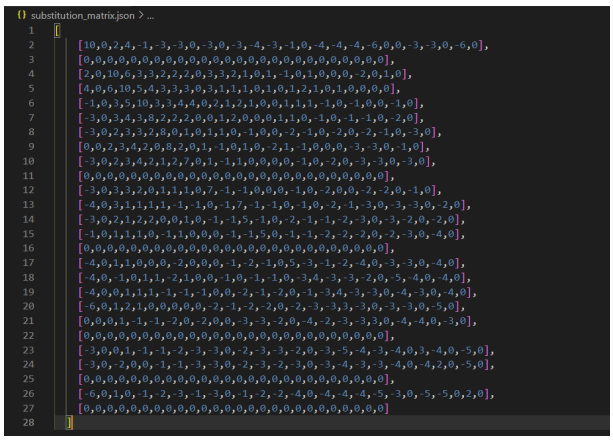

Figure 2: Substitution Matrix

2.1.3 GenomeAligner. To accomplish the alignment of the genome sequences and calculation of alignment scores, I adapted and utilized a substantial amount of code from CS-415. Specifically, I utilized the GenomeAligner Class, which includes methods for printing a matrix, identifying the best match, computing genome alignment scores, and determining the best alignment score from the previous pair of aligned genome sequences.

2.1.4 Utils. To perform genome sequence alignment and provide alignment scores, I developed a module called Utils.py. This module includes a wrapper method called perform_alignment_and_scoring(), which utilizes methods from the GenomeAligner class. Specifically, this wrapper method utilizes the calculate_alignment_score() method and the print_best_alignment_score_from_previous_pair() method to calculate the alignment scores from the provided substitution matrix and blossum50 matrix.

3 RESULTS

Below are the results generated by this project. These values demonstrate the range of variability in the aligned genome sequences. The use of the substitution matrix and Blossum50 matrix allowed for the alignment and scoring of the genome sequences, providing insight into their potential functional relationships and evolutionary history.

3.1 Blossum50 Global Alignment Example

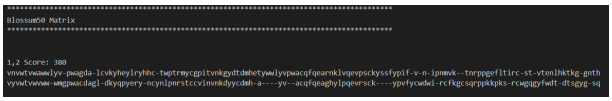

Figure 3: Blossum50 Global Alignment

3.2 Blossum50 Matrix Scores

	0	1	2	3	4	5	6	7
0	250	380	255	468	166	198	197	220
1	380	250	322	306	166	177	158	200
2	255	322	250	257	155	141	155	170
3	468	306	257	250	166	181	163	161
4	166	166	155	166	250	366	264	258
5	198	177	141	181	366	250	490	436
6	197	158	155	163	264	490	250	572
7	220	200	170	161	258	436	572	250

3.3 Blossum50 Matrix Min & Max Table

Input Sequence	Minimum	Maximum
0	166	468
1	158	380
2	141	322
3	161	468
4	155	366
5	141	490
6	155	572
7	161	572

Table 1: Blossum50 Matrix Minimum and Maximum Alignment Scores

3.4 Substitution Matrix Global Alignment Example

Figure 4: Substitution Matrix Global Alignment

3.5 Substitution Matrix Scores

	0	1	2	3	4	5	6	7
0	610	338	320	388	175	224	217	205
1	338	610	368	294	176	250	204	216
2	320	368	610	334	206	245	242	252
3	338	294	334	610	194	209	201	198
4	175	176	206	194	610	286	213	219
5	224	250	245	209	286	610	411	366
6	217	204	243	201	213	411	610	452
7	205	216	252	198	219	366	452	610

3.6 Substitution Matrix Min & Max Table

Input Sequence	Minimum	Maximum
0	175	610
1	176	610
2	206	610
3	194	610
4	175	610
5	209	610
6	201	610
7	198	610

Table 2: Substitution Matrix Minimum and Maximum Alignment Scores

4 CONCLUSION

In conclusion, the project successfully employed computational biology techniques, as presented in CS-415, to align a set of genome sequences and evaluate the alignment scores. The approach involved utilizing a substitution matrix generated in a previous project, in conjunction with the Blossum50 matrix. The results achieved using the methods employed in this study were deemed satisfactory. However, the confidence level in the generated outcomes is moderate. The results obtained from the two matrices were found to be dissimilar. Based on the analysis, it appears that the substitution matrix performed better overall, as it yielded higher alignment scores. However, it is possible that the genome sequences could have been clustered into two or more subsets to obtain more conclusive results. Overall, this investigation demonstrates the potential for utilizing genome alignment techniques to explore and understand genetic sequences, paving the way for further research in the field of computational biology.