Results:

```
In [87]: runfile('C:/Users/Arjunsinh Harer/OneDrive/Desktop/Personal Development/RBIF-109/week5/longest_common_subsequence.py', wdir='C:/
Users/Arjunsinh Harer/OneDrive/Desktop/Personal Development/RBIF-109/week5')
The longest subsequence of CGGTCTTATCGCTATCATACAGGTGAATT & TTGAGCCGTGTTGTGTCAACTTCATAATTCCCTGA, is a 5 base-pair subsequence TCATA

The longest subsequence of GGTCAGGTACTAACGAGTTT & GTCAACCATACTATCTTGCCGAAATTTA, is a 5 base-pair subsequence TACTA

The longest subsequence of CTTCTGGCGCAGAAGGGCAATTTGGCG & TCCAAGCGGTACCGTA, is a 3 base-pair subsequence GCG

The longest subsequence of CCAGTAATTACAATTAGAT & AACATGACTTGACGGCGTGCA, is a 3 base-pair subsequence ACA

The longest subsequence of CTTACTTGTGAAGCTGAGG & GTGTGTGCACCGTATGCGTGCAACCAAAGCGCCGGATTT, is a 4 base-pair subsequence TGTG

The longest subsequence of ACAGGAACTTCCCTGCGCAAAAACGGCAC & AGAACGTAGCGGATGGGACAAGCCTT, is a 4 base-pair subsequence GAAC

The longest subsequence of GCGTCAATCTCGGATTA & AATATAAGTACCACTGACGGACT, is a 4 base-pair subsequence CGGA

The longest subsequence of CGCGTTAGGAACTGATGCCCTAGTCG & GGGTGCGCGGGTATCGACGACACCTTATAGT, is a 4 base-pair subsequence CGCG

The longest subsequence of CGGCAGATTTGATCGATG & GTGAGGCTCGAAGATAGCTTTAGTAAGATTTATTTGAGAG, is a 6 base-pair subsequence AGATTT

The longest subsequence of AAGGGATTTCCCGTAGGAACGTTAACGTGAAACATGG & GGCCCTTTCTCGAACTTGACACGTAGGCGTGCCTAGGTT, is a 6 base-pair subsequence
CGTAGG
```

Fig 2. Output of the implementation, for the sake of readability, and easy visual validation I only created random base-pairs between 15-40 base pairs in length. But in theory this implementation can work for any length of DNA sequence.

Time complexity is O(n*m), where n & m are the length of the reference DNA sequences.

Conclusions:

This algorithm works well for short sequences, but in areas where the reference sequences exceed more than 10k kilobases the algorithm works slower. Additionally, this algorithm holds for two reference DNA sequences. But what if you wanted to find the longest common subsequence for 3 or more DNA sequences, the constraints increase. Extending this algorithm to multiple sequences does not scale because there will be exponentially more comparisons required. Additionally, if there are ties in the sequences the algorithm will only return one match. So even though the longest subsequence might be 4 base-pairs long, what if there are multiple 4 base-pair subsequences. Nonetheless, for comparing short reads between two reference DNA sequences, this algorithm operates effectively.

References:

[1]

"What is data deduplication? | Definition from TechTarget," Storage. Accessed: Feb. 13, 2024. [Online]. Available: https://www.techtarget.com/searchstorage/definition/data-deduplication

[2]

M. Goodrich and R. Tamassia, "Algorithm Design and Applications," Oct. 2014. Accessed: Feb. 13, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Algorithm-Design-and-Applications-Goodrich-Tamassia/6bf3b36fdf0430a109b1969562cc1449b0c0ad50

[3]

"Longest Common Substring | DP-29," GeeksforGeeks. Accessed: Feb. 13, 2024. [Online]. Available: https://www.geeksforgeeks.org/longest-common-substring-dp-29/