Distance Between Pattern and Strings

Input: A string *Pattern* followed by a collection of strings *Dna*
Output: $d(Pattern, Dna)$

SAMPLE DATASET:

Input:

AAA

TTACCTTAAC GATATCTGTC ACGGCGTTCG CCCTAAAGAG CGTCAGAGGT

Output:

5

The sample dataset is not actually run on your code.

TEST DATASET 1:
Input:
TAA
TTTATTT CCTACAC GGTAGAG
Output:
3

This dataset checks multiple potential mistakes.

First, it checks that you are actually using all three sequences of Dna (and not just a single sequence). The Hamming Distance between Pattern and each individual sequence in Dna is 1, so if your code returns a total score of 1, we fail it for this reason.

Next, it checks if you are only using the first k-mer in each sequence of Dna. For example, if you do this, you would output d(TAA,TTT) + d(TAA,CCT) + d(TAA,GGT) which is 8, instead of the correct answer of 3.

Finally, it checks if you are only using the last k-mer in each sequence of Dna. For example, if you do this, you would output d(TAA,TTT) + d(TAA,CAC) + d(TAA,GAG) which is 6, instead of the correct answer of 3.

TEST DATASET 2:

Input:

AAA

AAACT AAAC AAAG

Output:

0

This dataset checks if your code is using “max” or “sum” instead of “min”.

First, it checks if your code is using “max” instead of “min”. In this case, the output would be d(AAA,ACT) + d(AAA,AAC) + d(AAA,AAG), which is 4, instead of the correct answer of 0.

Next, it checks if your code is using “sum” instead of “min”. In this case, the output would be d(AAA,AAA) + d(AAA,AAC) + d(AAA,ACT) + d(AAA,AAA) + d(AAA,AAC) + d(AAA,AAA) + d(AAA,AAG), which is 5, instead of the correct answer of 0.

TEST DATASET 3:
Input:
AAA
TTTTAAA CCCCAAA GGGGAAA
Output:
0

This dataset checks if your code has an off-by-one error at the end of each sequence of Dna. Notice that each sequence has a perfect match of "AAA" at the very end, so if your code returns a nonzero answer to this test dataset, it must have missed the last k-mer of each.

TEST DATASET 4:
Input:
AAA
AAATTTT AAACCCC AAAGGGG
Output:
0

This dataset checks if your code has an off-by-one error at the beginning of each sequence of Dna. Notice that each sequence has a perfect match of "AAA" at the very beginning, so if your code returns a nonzero answer to this test dataset, it must have missed the first k-mer of each.