Application #4 Solution

Honor Code

To aid you in assessing your peers' submissions for the Application, we have developed an extensive solution/grading guide provided below that you should review **after** completing your first submission to the Application and before you assess your peers' submissions. **Please do not consult this reading before making your first attempt at the Application.** Using this reading to prepare your first submission violates the class Honor Code and robs you of the learning experience provided by this assignment. If you need to make a second submission for the Application, we ask that you limit your consultation of this solution guide and under no circumstances should you submit copies of the plots provided below. **Create your own original plots.**

Solution for Question 1 (2 pts)

Item a (1 pt) Is the score of the local alignment correct? (Hint: The sum of the decimal digits in the score is 20.)

The score for the local alignment is 875.

Item b (1 pt) Are the two sequences in the local alignments (with dashes included if inserted by the algorithm) clearly distinguished and correct?

The local alignment has the following two sequences:

The sequence for the HumanEyelessProtein is:

```
HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVV
  SKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEK-QQ
```

The sequence for the FruitflyEyelessProtein is:

```
HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVV
  SKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQ
```

Note that the local alignment is unique and the two submitted sequences should match exactly. Submitted answers that do not include the '-' in the human alignment should be counted as incorrect. You may wish to copy and paste the answer above in CodeSkulptor or desktop Python to do an exact comparison.

Solution for Question 2 (2 pts)

Item a (2 pts) Your answer should be two percentages: one for each sequence in the local alignment (human and fruitfly) computed in Question 1. Enter each percentage below. Be sure to clearly label each answer and include three significant digits of precision.

When globally-aligned, the HumanEyelessProtein sequence from the local alignment agrees with 72.9% of the ConsensusPAXDomain. The FruitflyEyelessProtein sequence from the local alignment agrees with 70.1% of the ConsensusPAXDomain when the two are globally aligned.

For scoring purposes, count any answer for the human/consensus that is between 72% and 73% as being correct, and count any answer for the fruitfly/consensus that is between 70% and 71% as being correct. Answers in decimal form are also acceptable.

Solution for Question 3 (1 pt)

Item a (1 pt) Examine your answers to Questions 1 and 2. Is it likely that the level of similarity exhibited by the answers could have been due to chance? Include a short informal justification for your answer.

The level of agreement of each local alignment to the ConsensusPAXDomain is much too high for this situation to have arisen by chance. Each alignment agrees on 90+ amino acids for sequences of length 130+. The chance that this many elements in each local alignment would agree with ConsensusPAXDomain due to chance would be vanishingly small for an alphabet of size 23, especially given so few dashes.

For the local human vs. consensus PAX domain alignment, the chance of 97 corresponding elements matching between two 133 element sequences whose elements are chosen at random from a 23 character alphabet is less than 10^{-100}!.

In evaluating the plausibility of the provided justification, examine the answer for a mention of the lengths of the sequences, the size of the alphabet, and the level of agreement. An estimate of the actual probability of agreement is not necessary. However, the answer should state the chance of agreement due to random matching is very, very small. If in doubt on this item, be generous.

Solution for Question 4 (2 pts)

Item a (1 pt) Does the plot follow the formatting guidelines for plots?

The formatting guidelines include the following items:

- The plot is an image and not a text file.
- The plot is appropriately trimmed. Showing the boundary of the plot's window is fine. However, the plot should not include part of the desktop.
- The elements of the plot are of the correct type. Line plots are not the same as point plots.
- Both axes should have tick marks labeled by regularly-spaced coordinate values.
- Both axes have appropriate text labels that describe the quantities being plotted.
- The plot has an appropriate title that describes the content of the plot.
- The plot has an appropriate legend (when required) that distinguishes the various components of the plot.

Assess the submitted plot based on these guidelines. Note that the submitted plot should be a bar plot that represents a normalized (vertical range between 0 and 1) distribution. If not, give no credit for this item.

Item b (1 pt) Is the shape of the plot correct?

Below is a bar plot created using 1000 trials with `generate_null_distribution`. Note that there will be some variation due to the use of random trials.

Image: http://storage.googleapis.com/codeskulptor-alg/alg_null_hypothesis_1000.png

When scoring the submitted plot, check for plots that are roughly bell-shaped in a manner similar to that of a normal distribution, but slightly asymmetric. The left-hand side of the distribution should drop to zero much more quickly than the right-hand side of the distribution, which trails off towards zero more slowly. The vast majority of the distribution should span a range from approximately 35 to 90 with a peak roughly centered at 50.

If the plot appears to have used substantially fewer trials (as indicated by a very noisy distribution), you may score this item as incorrect at your discretion.

Solution for Question 5 (2 pts)

Item a (1 pt) What are the mean and standard deviation for the distribution that you computed in Question 4?

The mean for our distribution in Question 4 is approximately 52. The standard deviation is approximately 6.8. Since the distribution may vary due to the use of random trials, score any mean in the range $[50, 55]$ as correct and any standard deviation in the range $[5.7, 8.0]$ as being correct. (These are very generous bounds.)

Item b (1 pt) What is the z-score for the local alignment for the human eyeless protein vs. the fruitfly eyeless protein based on these values?

For our computed distribution, the z-score was approximately 122. (Remember that s was 875.) Since this z-score may vary some due to the use of random trials, please score any z-score in the range $[100, 155]$ as being correct.

Solution for Question 6 (1 pt)

Item a (1 pt) Based on your answers to Questions 4 and 5, is the score resulting from the local alignment of the HumanEyelessProtein and the FruitflyEyelessProtein due to chance? As a concrete question, which is more likely: the similarity between the human eyeless protein and the fruitfly eyeless protein being due to chance or winning the jackpot in an extremely large lottery?

The distribution of scores is close enough to being bell-shaped that we will assume that 99% of the scores are withing three standard deviations of the mean for this distribution. Based on the z-score, the actual score for the Human/Fruitfly alignment is actually more than 100 standard deviations away from the mean of the distribution. If we assume that each multiple of three standard deviations reduces the likelihood of this score arising randomly by a factor of 10^{-2}, the resulting probability is on the order of approximately 10^{-67}. (In reality, the probability is much, much smaller.)

As a comparison, the odds of winning even the largest lottery are certainly more than one in a trillion (i.e; 10^{-12}). So, winning the jackpot in the world's largest lottery is much more likely.

When assessing the submitted answer check that the submitted answer mentions the fact that the score distribution has approximately bell-shaped (looks like a normal distribution) and that the z-score indicates that the true score is many standard deviations away from the mean.

Solution for Question 7 (3 pts)

Item a (3 pts) Determine the values for `diag_score`, `off_diag_score`, and `dash_score` such that the score from the resulting global alignment can be used to compute the edit distance via the formula above.

The correct values for the three types of entries in the scoring matrix are:

- `diag_score` is exactly 2,
- `off_diag_score` is exactly 1,
- `dash_score` is exactly 0.

The key to understanding the correctness of these values is to score the global alignments produced by this distance matrix on an character-by-character basis and compare this score to $|x| + |y|$. If two corresponding non-dash characters agree, the scoring matrix scores that match as 2. Note that these two matching characters also increase the size of $|x| + |y|$ by exactly two, leading to no increase in the edit distance.

If two corresponding non-dash characters disagree, the scoring matrix scores the match as 1. Since these two non-matching characters also increase the size of $|x| + |y|$ by exactly two, the edit distance is increased by one corresponding to the fact that a substitution is necessary. Finally, if a non-dashed character matches a dash, the scoring matrix scores this match as 0. Since the single non-dash character increases the size of size of $|x| + |y|$ by exactly one, the edit distance is increased by one corresponding to the fact that an insertion or deletion is necessary.

Solution for Question 8 (2 pts)

Item a (2 pts) Use your function `check_spelling` to compute the set of words with an edit distance of one from the string `"humble"` and the set of words with an edit distance of two from the string `"firefly"`. (Note this is not `"fruitfly"`.)

The set of words within edit distance one from the string `"humble"` is

```
set(['bumble', 'fumble', 'humble', 'humbled', 'humbler', 'humbles', 'humbly',
  'jumble', 'mumble', 'rumble', 'tumble'])
```

The set of words within an edit distance of two for the string `"firefly"` is

```
set(['direly', 'finely', 'fireclay', 'firefly', 'firmly', 'firstly', 'fixedly',
  'freely', 'liefly', 'refly', 'tiredly'])
```

When evaluating this item, each submitted word should exactly correspond to one of the words in the solution set. However, note the ordering of the submitted words is unimportant. The submitted answer does not have to appear in a particular format such as a set in Python.

Solution for Question 9 (optional, no credit)

Item a (1 pt) Reconsider the formulation of question 8 from a more general point of view and design a spelling correction tool that would provide real-time (almost instantaneous) correction of spelling errors. To guide you in the correct direction, we will provide two hints. First, you should convert your list of provided words to a set of words to enable a fast check for whether a string is a valid word. Second, you do not need to use dynamic programming to solve this problem. However, you will need to focus on the structure of the three editing operations describes in Question 7.

The key idea behind this improved spell checker is that we can detect whether a particular string is a valid word in $O(1)$ time by representing the word list as a set. Now, given a string that needs to be spell checked, we can first test whether the string is a valid word quickly. If so, we are done. If the string is not a valid word, we can use the editing operations in question 7 to generate all possible strings that are one edit away from the given string. Then, we can iterate through these strings to check whether any of these strings are valid words. This approach is also fast since the number of possible strings that are one edit away is still relatively small.

For words that are two edits away, we take the input string and enumerate all strings that are one edit away. Then, we take those strings and enumerate all strings that are one edit away from those strings. The result is the set of strings which are two edits away. We can then check whether those strings are in the provided word list. This page gives more details on how this approach is used by Google to do spelling correction very efficiently. The algorithm that Google uses also accounts for the fact that some words are used more often than others.

In scoring this question, the key observation to check for is whether the answer suggested building the set of all strings that are two edits away from the given string. If this observation is present in the explanation, give credit. If another method is proposed, use your judgment in deciding whether to award credit. Please add a comment describing your analysis of the proposed solution.

Mark as completed