ICS4U

ADT Assignment

DNA contains the blueprint to construct every living creature. It is made with only 4 base chemicals: adenine (A), guanine(G), cytosine(C) and thymine(T). Sequences of these 4 base chemicals form a gene specify how to make a single protein. For example, AATCCGTTCACC can be a sequence that describes some specific gene while the sequence AATGCCGGGATTC could be the sequence for describing a different gene. Sometimes a mutation could occur which may alter the sequence and give rise to a known disease. Other times the mutation may transform the original gene into another "valid" gene. Given a list of "valid" genes and "disease " genes for a particular species and the following "rules of mutation" can you determine the probability that a particular gene, P, can mutate into another gene, Q, within a certain number of mutations M?

Rules of Mutation- At each step one of the following can be done:

1) The starting base chemical can be swapped with the ending base chemical. This mutation has a 0.02% probability of occurring.

   Example: AGCCGT can mutate into TGCCGA

2) Any two identical, adjacent base chemicals can be replaced with a single base chemical. This mutation has a 0.06% probability of occurring.

   Example: TCCGA can mutate into TAGA

3) If the two base pairs G and T occur side by side in any order, another base chemical can be inserted in between them as long as the overall length of the gene does not exceed L. This mutation has a 0.08% probability of occurring.

   Example: TCCTGA can mutate into TCCTCGA

INPUT: a file called DATA.TXT in the root of your project that contains the following:

The first line contains the maximum length L of any gene.
The second line contains an integer, V, indicating the number of valid genes.
The third line contains an integer, D, indicating the number of known disease genes.
The next V lines each contain one valid gene of maximum size L;
The next D lines each contain one disease gene of maximum size L;
The next line contains an integer M indicating the maximum number of allowable mutations
The next line contains an integer G indicating the number of genes to test.
Followed by G lines contain the initial gene P, followed by the mutated gene Q.

The limits on the inputs are as follows:

$1 < L <= 10$
$1 < V + D <= 20000$
$1 < M <= 10$
$1 < G <= 5$

OUTPUT:

For each of the G test cases, the output consists of two lines:
1) Either YES or NO depending on whether gene P can mutate into gene Q within M mutation steps.
2) If the answer to 1) was YES, then the second line outputs the largest probability that P can mutate into a gene Q within the given steps, otherwise nothing is printed.

ICS4U

EXAMPLE

INPUT:

```
4
8
2
AGT
AGG
AC
GTT
GGC
GAC
GGA
CAG
GATT
TGG
4
3
GTT  CAG
GGA  AC
AGT  TGG
```

OUTPUT:

```
YES
9.599999999999999E-5
YES
0.012
NO
```

What To Hand In

1. Your Java based solution to the above problem, including proper header files and comments as usual. If your submission includes more than one file, please create an archive of your project using Eclipse and submit it.
2. A written description of how your algorithm works. Make sure to use the concepts described in class. Your ADT(s) must be entirely written by you; you cannot use any predefined Java classes.
3. A worst-case run time analysis of your algorithm in Big-O notation.