The Genetic Code

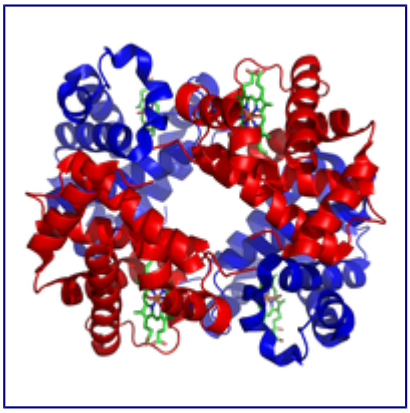

Figure 1. The human hemoglobin molecule consists of 4 polypeptide chains; α subunits are shown in red and β subunits are shown in blue

Just as nucleic acids are polymers of nucleotides, proteins are chains of smaller molecules called amino acids; 20 amino acids commonly appear in every species. Just as the primary structure of a nucleic acid is given by the order of its nucleotides, the primary structure of a protein is the order of its amino acids. Some proteins are composed of several subchains called polypeptides, while others are formed of a single polypeptide; see Figure 1.

Proteins power every practical function carried out by the cell, and so presumably, the key to understanding life lies in interpreting the relationship between a chain of amino acids and the function of the protein that this chain of amino acids eventually constructs. Proteomics is the field devoted to the study of proteins.

How are proteins created? The genetic code, discovered throughout the course of a number of ingenious experiments in the late 1950s, details the translation of an RNA molecule called messenger RNA (mRNA) into amino acids for protein creation. The apparent difficulty in translation is that somehow 4 RNA bases must be translated into a language of 20 amino acids; in order for every possible amino acid to be created, we must translate 3-nucleobase strings (called codons) into amino acids. Note that there are 4^3=64 possible codons, so that multiple codons may encode the same amino acid. Two special types of codons are the start codon (AUG), which codes for the amino acid methionine always indicates the start of translation, and the three stop codons (UAA, UAG, UGA), which do not code for an amino acid and cause translation to end.

The notion that protein is always created from RNA, which in turn is always created from DNA, forms the central dogma of molecular biology. Like all dogmas, it does not always hold; however, it offers an excellent approximation of the truth.

An organelle called a ribosome creates peptides by using a helper molecule called transfer RNA (tRNA). A single tRNA molecule possesses a string of three RNA nucleotides on one end (called an anticodon) and an amino acid at the other end. The ribosome takes an RNA molecule transcribed from DNA (see "Transcribing DNA into RNA"), called messenger RNA (mRNA), and examines it one codon at a time. At each step, the tRNA possessing the complementary anticodon bonds to the mRNA at this location, and the amino acid found on the opposite end of the tRNA is added to the growing peptide chain before the remaining part of the tRNA is ejected into the cell, and the ribosome looks for the next tRNA molecule.

Not every RNA base eventually becomes translated into a protein, and so an interval of RNA (or an interval of DNA translated into RNA) that does code for a protein is of great biological interest; such an interval of DNA or RNA is called a gene. Because protein creation drives cellular processes, genes differentiate organisms and serve as a basis for heredity, or the process by which traits are inherited.

Problem

The 20 commonly occurring amino acids are abbreviated by using 20 letters from the English alphabet (all letters except for B, J, O, U, X, and Z). Protein strings are constructed from these 20 symbols. Henceforth, the term genetic string will incorporate protein strings along with DNA strings and RNA strings.

The RNA codon table dictates the details regarding the encoding of specific codons into the amino acid alphabet.

Given: An RNA string *s* corresponding to a strand of mRNA (of length at most 10 kbp).

Return: The protein string encoded by *s*.

Sample Dataset

AUGGCCAUGGCGCCCAGAACUGAGAUCAAUAGUACCCGUAUUAACGGGUGA

Sample Output

```
MAMAPRTEINSTRING
```

Combing Through the Haystack

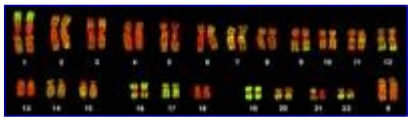

Figure 1. The human chromosomes stained with a probe for Alu elements, shown in green.

Finding the same interval of DNA in the genomes of two different organisms (often taken from different species) is highly suggestive that the interval has the same function in both organisms.

We define a motif as such a commonly shared interval of DNA. A common task in molecular biology is to search an organism's genome for a known motif.

The situation is complicated by the fact that genomes are riddled with intervals of DNA that occur multiple times (possibly with slight modifications), called repeats. These repeats occur far more often than would be dictated by random chance, indicating that genomes are anything but random and in fact illustrate that the language of DNA must be very powerful (compare with the frequent reuse of common words in any human language).

The most common repeat in humans is the Alu repeat, which is approximately 300 bp long and recurs around a million times throughout every human genome (see Figure 1). However, Alu has not been found to serve a positive purpose, and appears in fact to be parasitic: when a new Alu repeat is inserted into a genome, it frequently causes genetic disorders.

Problem

Given two strings s and t, t is a substring of s if t is contained as a contiguous collection of symbols in s (as a result, t must be no longer than s).

The position of a symbol in a string is the total number of symbols found to its left, including itself (e.g., the positions of all occurrences of 'U' in "AUGCUUCAGAAAGGUCUUACG" are 2, 5, 6, 15, 17, and 18). The symbol at position I of s is denoted by $s[i]$.

A substring of s can be represented as $s[j{:}k]$, where j and k represent the starting and ending positions of the substring in s; for example, if s = "AUGCUUCAGAAAGGUCUUACG", then $s[2{:}5]$ = "UGCU".

The location of a substring $s[j{:}k]$ is its beginning position j; note that t will have multiple locations in s if it occurs more than once as a substring of s (see the Sample below).

Given: Two DNA strings s and t (each of length at most 1 kbp).

Return: All locations of t as a substring of s.

Sample Dataset

```
GATATATGCATATACTT
ATAT
```

Sample Output

```
2 4 10
```