Searching Through the Haystack

> In “Finding a Motif in DNA”, we searched a given genetic string for a motif; however, this problem assumed that we know the motif in advance. In practice, biologists often do not know exactly what they are looking for. Rather, they must hunt through several different genomes at the same time to identify regions of similarity that may indicate genes shared by different organisms or species.
>
> The simplest such region of similarity is a motif occurring without mutation in every one of a collection of genetic strings taken from a database; such a motif corresponds to a substring shared by all the strings. We want to search for long shared substrings, as a longer motif will likely indicate a greater shared function.

Problem

A common substring of a collection of strings is a substring of every member of the collection. We say that a common substring is a longest common substring if there does not exist a longer common substring. For example, "CG" is a common substring of "A**CG**TACGT" and "AAC**CG**TATA", but it is not as long as possible; in this case, "CGTA" is a longest common substring of "A**CGTA**CGT" and "AAC**CGTA**TA".

Note that the longest common substring is not necessarily unique; for a simple example, "AA" and "CC" are both longest common substrings of "AACC" and "CCAA".

Given: A collection of $k (k \leq 100)$ DNA strings of length at most 1 kbp each in FASTA format.

Return: A longest common substring of the collection. (If multiple solutions exist, you may return any single solution.)

Sample Dataset

```
>Rosalind_1
GATTACA
>Rosalind_2
TAGACCA
>Rosalind_3
ATACA
```

Sample Output

```
AC
```

Motif Implies Functionclick to collapse

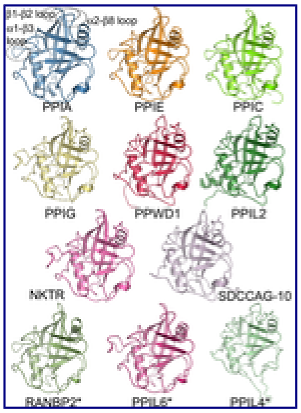

Figure 1. The human cyclophilin family, as represented by the structures of the isomerase domains of some of its members.

As mentioned in “Translating RNA into Protein”, proteins perform every practical function in the cell. A structural and functional unit of the protein is a protein domain: in terms of the protein's primary structure, the domain is an interval of amino acids that can evolve and function independently.

Each domain usually corresponds to a single function of the protein (e.g., binding the protein to DNA, creating or breaking specific chemical bonds, etc.). Some proteins, such as myoglobin and the Cytochrome complex, have only one domain, but many proteins are multifunctional and therefore possess several domains. It is even possible to artificially fuse different domains into a protein molecule with definite properties, creating a chimeric protein.

Just like species, proteins can evolve, forming homologous groups called protein families. Proteins from one family usually have the same set of domains, performing similar functions; see Figure 1.

A component of a domain essential for its function is called a motif, a term that in general has the same meaning as it does in nucleic acids, although many other terms are also used (blocks, signatures, fingerprints, etc.) Usually protein motifs are evolutionarily conservative, meaning that they appear without much change in different species.

Proteins are identified in different labs around the world and gathered into freely accessible databases. A central repository for protein data is UniProt, which provides detailed protein annotation, including function description, domain structure, and post-translational modifications. UniProt also supports protein similarity search, taxonomy analysis, and literature citations.

Problem

To allow for the presence of its varying forms, a protein motif is represented by a shorthand as follows: [XY] means "either X or Y" and {X} means "any amino acid except X." For example, the N-glycosylation motif is written as N{P}[ST]{P}.

We can see the complete description and features of a particular protein by its access ID "uniprot_id" in the UniProt database, by inserting the ID number into

```
http://www.uniprot.org/uniprot/uniprot_id
```

Alternatively, we can obtain a protein sequence in FASTA format by following

```
http://www.uniprot.org/uniprot/uniprot_id.fasta
```

For example, the data for protein B5ZC00 can be found at http://www.uniprot.org/uniprot/B5ZC00.

Given: At most 15 UniProt Protein Database access IDs.

Return: For each protein possessing the N-glycosylation motif, output its given access ID followed by a list of locations in the protein string where the motif can be found.

Sample Dataset

```
A2Z669
B5ZC00
P07204_TRBM_HUMAN
P20840_SAG1_YEAST
```

Sample Output

```
B5ZC00
85 118 142 306 395
P07204_TRBM_HUMAN
47 115 116 382 409
P20840_SAG1_YEAST
79 109 135 248 306 348 364 402 485 501 614
```

Genes are Discontiguous

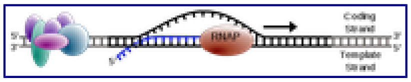

Figure 1. The elongation of a pre-mRNA by RNAP as it moves down the template strand of DNA.

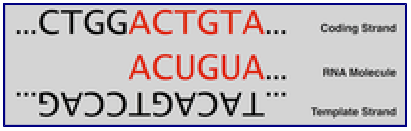

Figure 2. RNA is identical to the coding strand except for the replacement of thymine with uracil.

In “Transcribing DNA into RNA”, we mentioned that a strand of DNA is copied into a strand of RNA during transcription, but we neglected to mention how transcription is achieved.

In the nucleus, an enzyme (i.e., a molecule that accelerates a chemical reaction) called RNA polymerase (RNAP) initiates transcription by breaking the bonds joining complementary bases of DNA. It then creates a molecule called precursor mRNA, or pre-mRNA, by using one of the two strands of DNA as a template strand: moving down the template strand, when RNAP encounters the next nucleotide, it adds the complementary base to the growing RNA strand, with the provision that uracil must be used in place of thymine; see Figure 1.

Because RNA is constructed based on complementarity, the second strand of DNA, called the coding strand, is identical to the new strand of RNA except for the replacement of thymine with uracil. See Figure 2 and recall “Transcribing DNA into RNA”.

After RNAP has created several nucleotides of RNA, the first separated complementary DNA bases then bond back together. The overall effect is very similar to a pair of zippers traversing the DNA double helix, unzipping the two strands

> and then quickly zipping them back together while the strand of pre-mRNA is produced.
>
> For that matter, it is not the case that an entire substring of DNA is transcribed into RNA and then translated into a peptide one codon at a time. In reality, a pre-mRNA is first chopped into smaller segments called introns and exons; for the purposes of protein translation, the introns are thrown out, and the exons are glued together sequentially to produce a final strand of mRNA. This cutting and pasting process is called splicing, and it is facilitated by a collection of RNA and proteins called a spliceosome. The fact that the spliceosome is made of RNA and proteins despite regulating the splicing of RNA to create proteins is just one manifestation of a molecular chicken-and-egg scenario that has yet to be fully resolved.
>
> In terms of DNA, the exons deriving from a gene are collectively known as the gene's coding region.

Problem

After identifying the exons and introns of an RNA string, we only need to delete the introns and concatenate the exons to form a new string ready for translation.

Given: A DNA string s(of length at most 1 kbp) and a collection of substrings of s acting as introns. All strings are given in FASTA format.

Return: A protein string resulting from transcribing and translating the exons of s. (Note: Only one solution will exist for the dataset provided.)

Sample Dataset

```
>Rosalind_10
ATGGTCTACATAGCTGACAAACAGCACGTAGCAATCGGTCGAATCTCGAGAGGCATATGGTCACATGATCGGTCGAGCGTGT
TTCAAAGTTTGCGCCTAG
>Rosalind_12
ATCGGTCGAA
>Rosalind_15
ATCGGTCGAGCGTGT
```

Sample Output

```
MVYIADKQHVASREAYGHMFKVCA
```