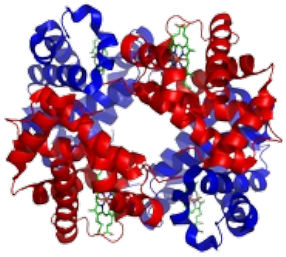
Infer mRNA from protein

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. 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. The field devoted to the study of proteins is called proteomics.



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

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 synthesis. 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 codons** (AUG in the standard genetic code), which code for the amino acid methionine and always indicate the start of translation, and the stop codons (UAA, UAG, UGA in the standard genetic code), 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. A eukaryotic 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 mRNA molecule transcribed from DNA 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.

Assignment

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. An **RNA codon table** dictates the details regarding the encoding of specific codons into the amino acid alphabet. The codon table shown below gives the mapping used by the standard genetic code. However, there are alternative genetic codes that show light variations on the mapping scheme. Stop codons do not code for an amino acid, and are indicated by the word Stop instead of an amino acid symbol.

```
UUU F      CUU L      AUU I      GUU V
UUC F      CUC L      AUC I      GUC V
UUA L      CUA L      AUA I      GUA V
UUG L      CUG L      AUG M      GUG V
UCU S      CCU P      ACU T      GCU A
UCC S      CCC P      ACC T      GCC A
UCA S      CCA P      ACA T      GCA A
UCG S      CCG P      ACG T      GCG A
UAU Y      CAU H      AAU N      GAU D
UAC Y      CAC H      AAC N      GAC D
UAA Stop   CAA Q      AAA K      GAA E
UAG Stop   CAG Q      AAG K      GAG E
UGU C      CGU R      AGU S      GGU G
UGC C      CGC R      AGC S      GGC G
UGA Stop   CGA R      AGA R      GGA G
UGG W      CGG R      AGG R      GGG G
```

When researchers discover a new protein, they would like to infer the strand of mRNA from which this protein could have been translated, thus allowing them to locate genes associated with this protein on the genome. Unfortunately, although any RNA string can be translated into a unique protein string, reversing the process yields a huge number of possible RNA strings from a single protein string because most amino acids correspond to multiple RNA codons. For a given protein string, determine the total number of different mRNA strings from which the protein could have been translated. In order to do this, you proceed as follows:

- Write a function codontable that takes the location of a text file as an argument. This text file must contain the mapping from codons to amino acids as used by a genetic code, in the format as shown in the table above. The function must read the mapping from the text file, and return it as a dictionary (that maps each of the 64 possible codons into their corresponding amino acid). Stop codons should be mapped onto an asterisk (*).
- Write a function reverse_codontable that takes a dictionary having the form of the dictionaries returned by the function codontable. The function must return a new dictionary that maps each of the 20 amino acids and the stop codon (represented by an asterisk: *) onto the set of codons that translate to this amino acid/stopcodon.
- Write a function mRNA that takes two arguments: a protein string and a codon table. The codon table must be passed as a dictionary having the form of the dictionaries returned by the function codontable. The function must return the total number of different mRNA strings from which the protein could have been translated. This number can be written as the product of the number of codons that gets mapped onto each amino acid in the protein string. Take into account that protein synthesis ends at a stop codon (which is not explicitly indicated at the end of the protein string that is passed to the function). This means you also have to multiply by the number of stop codons. An example: according to the codon table of the standard genetic code there 12 mRNA strings that translate to the protein MA, because there is a single codon translating to M, four to A and there are three stop codons: $1

\times 4 \times 3 = 12$. Along the same way, there are $1 \times 6 \times 2 \times 4 \times 3 = 144$ mRNA strings that translate to the protein string MRNA.

Example

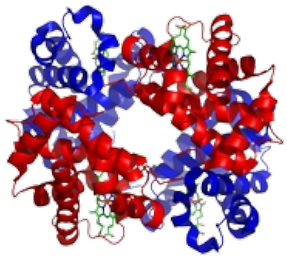
In the following interactive session we assume that the text file standard_code.txt is located in the current directory.

```
>>> table = codontable('standard_code.txt')
>>> table
{'GUC': 'V', 'ACC': 'T', 'GUA': 'V', 'GUG': 'V', 'GUU': 'V', 'AAC': 'N', 'CCU': 'P', 'UGG': 'W', 'AGC': 'S', 'AUC': 'I', 'CAU': 'H', 'AAU': 'N', 'AGU': 'S', 'ACU': 'T', 'CAC': 'H', 'ACG': 'T', 'CCG': 'P', 'CCA': 'P', 'AC

>>> reverse_codontable(table)
{'A': {'GCA', 'GCC', 'GCU', 'GCG'}, 'C': {'UGC', 'UGU'}, 'E': {'GAG', 'GAA'}, 'D': {'GAU', 'GAC'}, 'G': {'GGU', 'GGG', 'GGA', 'GGC'}, 'F': {'UUU', 'UUC'}, 'I': {'AUA', 'AUC', 'AUU'}, 'H': {'CAC', 'CAU'}, 'K': {'AA

>>> mRNA('MA', table)
12
>>> mRNA('MWQWQWY', table)
24
>>> mRNA('MRNA', table)
144
```

Net zoals nucleïnezuren polymeren zijn van nucleotiden, zijn **eiwitten** ketens van kleinere moleculen die **aminozuren** genoemd worden. In elk levend organisme vinden we de gebruikelijke 20 aminozuren. Net zoals de primaire structuur van een nucleïnezuur bepaald wordt door de volgorde van zijn nucleotiden, wordt de primaire structuur van een eiwit bepaald door de volgorde van zijn aminozuren. Sommige eiwitten bestaan uit verschillende deelketens die polypeptiden genoemd worden, terwijl andere uit slechts één enkele polypeptide bestaan. Eiwitten zijn de motor van elke functie die door de cel wordt uitgevoerd, en dus ligt de sleutel tot het ontsluieren van leven vermoedelijk in het interpreteren van de relatie tussen een keten van aminozuren en de functie van het eiwit dat uiteindelijk gevormd wordt door deze keten van aminozuren. Het onderzoeksdomein dat gewijd is aan de studie van eiwitten wordt *proteomics* genoemd.



Het menselijk hemoglobinemolecuul bestaat uit vier polypetideketens; α subeenheden zijn aangegeven in het rood en β subeenheden zijn aangegeven in het blauw.

Hoe worden eiwitten gemaakt? De **genetische code** — ontdekt doorheen een aantal ingenieuze experimenten die in de late jaren '50 uitgevoerd werden — geeft in detail weer hoe de vertaling verloopt van een RNA molecule die **boodschapper RNA** (messenger RNA, mRNA) genoemd wordt naar de aminozuren die instaan voor eiwitsynthese. De ogenschijnlijke moeilijkheid bij deze vertaling is dat ze op één of andere manier in staat moet zijn om 4 RNA basen te vertalen naar een taal die bestaat uit 20 aminozuren. Om elk van de mogelijke aminozuren te kunnen maken, moeten we 3-nucleobase strings (die in deze context **codons** genoemd worden) vertalen naar aminozuren. Merk op dat er $4^3=64$ mogelijke codons zijn, waardoor verschillende codons kunnen coderen voor hetzelfde aminozuur. Twee speciale soorten codons zijn de **startcodons** (AUG in de standaard genetische code) die coderen voor het aminozuur methionine en altijd aangeven waar de vertaling begint, en de **stopcodons** (UAA, UAG, UGA in de standaard genetische code), die niet coderen voor een aminozuur maar aangeven waar de vertaling van het eiwit eindigt.

De idee dat eiwitten altijd gemaakt worden uit RNA, dat op zijn beurt altijd gemaakt wordt uit DNA, vormt het centrale dogma van de moleculaire biologie. Net als alle andere dogma's is het niet altijd waar. Het biedt ons echter een uitstekende benadering van de werkelijkheid. Het ribosoom maakt peptiden aan met behulp van een hulpmolecule die **transfer RNA** (tRNA) genoemd wordt. Een tRNA molecule heeft een reeks van drie RNA nucleotiden aan één uiteinde (dit wordt een anticodon genoemd) en een aminozuur aan het andere uiteinde. Het ribosoom neemt een mRNA molecule die aangemaakt werd uit DNA en onderzoekt het codon per codon. Bij elke stap bindt een tRNA molecule die beschikt over het complementaire anticodon zich aan het codon op die positie, en wordt het aminozuur aan het andere uiteinde van het tRNA toegevoegd aan de groeiende peptideketen vooraleer het resterende deel van het tRNA wordt uitgestoten in de cel. Daarna gaat het ribosoom op zoek naar de volgende tRNA molecule.

Niet elke RNA base zal uiteindelijk vertaald worden naar een eiwit, waardoor een interval van RNA (of een interval van DNA dat vertaald wordt naar RNA) dat codeert voor een eiwit biologisch gezien een zeer belangrijke waarde heeft. Een dergelijk interval van DNA of RNA wordt een gen genoemd. Omdat eiwitsynthese de drijvende kracht is achter alle cellulaire processen, zorgen genen voor het onderscheid tussen organismen en vormen ze de basis voor erfelijkheid, zijnde het proces waarbij eigenschappen worden overgeërfd.

Opgave

De 20 gebruikelijke aminozuren worden symbolisch voorgesteld door 20 hoofdletters van ons alfabet (alle letters behalve B, J, O, U, X, en Z). Eiwitstrings worden opgebouwd uit deze 20 letters. Een **RNA codontabel** bevat alle details omtrent het coderen van specifieke codons naar het alfabet van de aminozuren. De codontabel die hieronder staat, geeft de afbeelding die gebruikt wordt bij de standaard genetische code. Er bestaan echter alternatieve genetische codes die kleine verschillen vertonen met dit afbeeldingsschema. Stopcodons coderen niet voor een aminozuur en worden in de tabel aangegeven met het woord Stop in plaats van een letter die staat voor een aminozuur.

```
UUU F      CUU L      AUU I      GUU V
UUC F      CUC L      AUC I      GUC V
UUA L      CUA L      AUA I      GUA V
UUG L      CUG L      AUG M      GUG V
UCU S      CCU P      ACU T      GCU A
UCC S      CCC P      ACC T      GCC A
UCA S      CCA P      ACA T      GCA A
UCG S      CCG P      ACG T      GCG A
UAU Y      CAU H      AAU N      GAU D
UAC Y      CAC H      AAC N      GAC D
```

UAA Stop CAA Q AAA K GAA E
UAG Stop CAG Q AAG K GAG E
UGU C CGU R AGU S GGU G
UGC C CGC R AGC S GGC G
UGA Stop CGA R AGA R GGA G
UGG W CGG R AGG R GGG G

Als onderzoekers een nieuw eiwit ontdekken, dan willen ze graag achterhalen van welke RNA streng dit eiwit kan vertaald zijn, om zo de genen die geassocieerd zijn met dit eiwit te kunnen localiseren op het genoom. Ondanks het feit dat elke RNA string op een unieke manier kan vertaald worden naar een eiwitstring, levert het omgekeerde proces helaas een groot aantal mogelijke RNA strings op voor eenzelfde eiwitstring omdat aminozuren mogelijks corresponderen met meerdere RNA codons. Je opdracht bestaat erin om voor een gegeven eiwitstring te bepalen hoeveel verschillende mRNA strings een vertaling geven naar dit eiwit. Hiervoor ga je als volgt te werk:

- Schrijf een functie `codontabel` waaraan de locatie van een tekstbestand moet doorgegeven worden. Dit tekstbestand moet de afbeelding van codons naar aminozuren bevatten die gebruikt wordt door een bepaalde genetische code, in het formaat zoals dat hierboven wordt weergegeven. De functie moet de afbeeldint uit het bestand lezen, en teruggeven onder de vorm van een dictionary (die dus de 64 mogelijke codons afbeeldt op hun corresponderende aminozuur). Stopcodons moeten hierbij afgebeeld worden op een sterretje (`*`).
- Schrijf een functie `omgekeerde_codontabel` waaraan een dictionary moet doorgegeven worden die de vorm heeft van de dictionaries die teruggegeven worden door de functie `codontabel`. De functie moet een nieuwe dictionary teruggeven, die elk van de 20 aminozuren + het stopcodon (voorgesteld door een sterretje: `*`) afbeeldt op de verzameling van codons die vertalen naar dit aminozuur/stopcodon.
- Schrijf een functie `mRNA` waaraan twee argumenten moeten doorgegeven worden: een eiwitstring en een codontabel. De codontabel moet doorgegeven worden als een dictionary die de vorm heeft van de dictionaries die teruggegeven worden door de functie `codontabel`. De functie moet teruggeven hoeveel verschillende mRNA strings volgens de gegeven codontabel vertaald worden naar het gegeven eiwit. Dit aantal kan bepaald worden als het product van het aantal codons dat afgebeeld wordt op elk van de aminozuren in de eiwitstring. Vergeet ook niet dat eiwitsynthese eindigt bij een stopcodon (dat niet expliciet is aangegeven op het einde van de eiwitstring die aan de functie wordt doorgegeven). Er moet dus ook vermenigvuldigd worden met het aantal mogelijke stopcodons. Een voorbeeld: volgens de codontabel van de standaard genetische code worden bijvoorbeeld 12 mRNA strings vertaald naar het eiwit `MA`, want er is één codon dat vertaald wordt naar `M`, vier naar `A` en er zijn ook drie stopcodons: $1 \times 4 \times 3 = 12$. Op analoge manier vertalen er $1 \times 6 \times 2 \times 4 \times 3 = 144$ mRNA strings naar de eiwitstring `MRNA`.

Voorbeeld

Bij onderstaande voorbeeldsessie gaan we ervan uit dat het tekstbestand standard_code.txt zich in de huidige directory bevindt.

```
>>> tabel = codontabel('standard_code.txt')
>>> tabel
{'GUC': 'V', 'ACC': 'T', 'GUA': 'V', 'GUG': 'V', 'GUU': 'V', 'AAC': 'N', 'CCU': 'P', 'UGG': 'W', 'AGC': 'S', 'AUC': 'I', 'CAU': 'H', 'AAU': 'N', 'AGU': 'S', 'ACU': 'T', 'CAC': 'H', 'ACG': 'T', 'CCG': 'P', 'CCA': 'P', 'AC

>>> omgekeerde_codontabel(tabel)
{'A': {'GCA', 'GCC', 'GCU', 'GCG'}, 'C': {'UGC', 'UGU'}, 'E': {'GAG', 'GAA'}, 'D': {'GAU', 'GAC'}, 'G': {'GGU', 'GGG', 'GGA', 'GGC'}, 'F': {'UUU', 'UUC'}, 'I': {'AUA', 'AUC', 'AUU'}, 'H': {'CAC', 'CAU'}, 'K': {'AA

>>> mRNA('MA', tabel)
12
>>> mRNA('MWQWQWY', tabel)
24
>>> mRNA('MRNA', tabel)
144
```