Some Basic Terminology

These are notes to provide basic terminology related to the bioinformatics modules in the computer science course sequence. This is a first draft; if you have suggestions, please send them to me.

Background

A *DNA string*, also called a *DNA strand,* is a finite sequence consisting of the four letters `A`, `C`, `G`, and `T` in any order[1]. The four letters stand for the four *nucleotides*: *adenine*, *cytosine*, *guanine*, and *thymine*. Nucleotides, which are the molecular units from which DNA and RNA are composed, are also called *bases*. Each nucleotide has a *complement* among the four: `A` and `T` are complements, and `C` and `G` are complements. Complements are chemically-related in that when they are close to each other, they form *hydrogen bonds* between them.

A special enzyme called *RNA polymerase* uses the information in DNA to create RNA. The process of creating RNA from DNA is called *transcription*. A *RNA string* or *RNA strand* is a finite sequence consisting of the four lowercase letters `A`, `C`, `G`, and `U`. The `A`, `C`, and `G` have the same names as they do in DNA, but the `U` represents *uracil*. When DNA is transcribed to RNA by RNA polymerase, each thymine base is converted to uracil. Hence RNA strings have `U`'s wherever DNA has `T`'s.

RNA in turn serves as a template for the construction of *proteins*, which are sequences of *amino acids*. Proteins are synthesized within the *ribosomes* of living cells by a process called *translation*. In translation, the RNA string is viewed as a sequence of three-letter groups called *codons*. Each codon codes for a particular amino acid. For example, `GUU` codes for *valine*, and `UCA` codes for *cysteine*. Let us count how many possible three-letter sequences are there in which each letter can be `A`, `C`, `G`, or `T`. There are four choices for the first letter, four independent choices for the second letter, and four for the third, so there are $4^3 = 64$ different codons. On the other hand, there are only 20 different amino acids. Some amino acids are coded for by multiple codons. For example, `UCA`, `UCC`, `UCG`, and `UCU` all code for *cysteine*. Some codons do not code for any amino acids; they are *stop codes*. There are three stop codons: `UAA`, `UAG`, and `UGA`.

Stop codes are used during protein synthesis to terminate reading of the RNA string. Not all of a RNA string is translated into protein; there are large regions that act like gaps. As the RNA is read, when a gap is reached, it is skipped over until a special start codon is found that tells the ribosome to begin creating amino acids again. When it sees a stop codon it stops and keeps reading until it finds another start codon, and so on, until the entire strand is read. The process is just like the way comments in Perl programs are treated by the Perl compiler. The `#` tells the compiler to stop compiling, and the next newline character tells it to start compiling again. The `#` is a stop code and the newline is a start code. As an example, the RNA strand

```
AUGGUUUAUGGUCUCUGA
```

is read as the following sequence of codons

```
AUG GUU UAU GGU CUC UGA
```

Consulting a table of these mappings, we see that `AUG` is a start codon that codes for *methionine* (Met), `GUU`, for *valine* (Val), `UAU`, for *tyrosine* (Tyr), `GGU`, for *glycine* (Gly), `CUC`, for *leucine* (Leu), and `UGA` is a stop codon. Therefore, the sequence *Met-Val-Tyr-Gly-Leu* is created from this RNA fragment.

[1] Some sources use lowercase while others use uppercase. Here they will be used interchangeably

Amino acids have long names like cysteine but they also have three-letter names such as Cys, for cysteine, and one-letter uppercase names, such as C for cysteine. It is not always true that the one-letter name is the first letter of the amino acid's long name. The above sequence would be written *MVYGL* using the one-letter names.

Direction and Shape

In its most common form, DNA is actually a double helix consisting of two strands that wrap around each other. Each DNA strand has *direction*. Direction is usually indicated by putting a 5' at one end and a 3' at the other. The 5' and 3' refer to the names of the carbon atoms to which these ends attach. The carbons are part of a ring structure with multiple carbon atoms, so they get names for the purpose of distinuishing them from each other. For example,

```
5'-GTATCC-3'
```

is a fragment of DNA that runs from the 5' to the 3' position. The 5' end is called the *upstream* end, and the 3' end is the *downstream* end. The two strands of nucleotides are in reverse directions of each other. In other words, if you could unwind the helix so that the two strands were lying on a flat surface parallel to each other, in one strand the 5' end would be to the left, and in the other, it would be to the right. The two strands are chemically-related because the bases that would be across from each other on the table are complements of each other. For example, the two strands below

```
5'-G T A T C C A A T G C C-3'
   | | | | | | | | | | | |
3'-C A T A G G T T A C G G-5'
```

could be a fragment of the unwound double helix. The vertical lines connect complements in the forward and reverse strands to each other. Each `C` in one is matched by a `G` in the other, and each `A` is matched by a `T` in the other.

Characteristic Properties of DNA

Scientists use various heuristic rules in their study of DNA. Among the many metrics that they use are the following:

- A *poly-T sequence* of length N is a sequence of N or more consecutive `T` nucleotides.
- The *GC content* of a DNA strand is the ratio of the total number of `C` and `G` nucleotides to the length of the strand. For example, the sequence '`atcgtttgga`' is of length 10 and has a total of 4 `C`'s and `G`'s, so its GC content is 0.4.
- A *CpG island* is a `C` followed by a `G` in a DNA strand. (The *p* in between the `C` and `G` represents the fact that a *phosphodiester* bond connects `them`.)

Restriction Enzymes

Bacteria produce special enzymes that can cut their DNA at specified sites, called *cleavage sites*. The cleavage site is a position between two nucleotides in the DNA. The enzyme finds its site by a type of biological pattern-matching. The pattern specifies where in the DNA the enzyme will match. For example, the enzyme *EcoRI* has a recognition site defined by

```
5'-G'AATTC-3'
```

This means that it will search for a substring of the DNA consisting of the bases `GAATTC` in the 5' to 3' direction, and cut the DNA between the `G` and the first `A`. The apostrophe ' indicates the cleavage site. So, if the DNA string is

```
ATGAAAGGGTTTCCCTTTGAATTCCCCATGGTATTGTTGCCGGAATTCTTTCCGGCCCCC
```

it will be cut into the three pieces

```
ATGAAAGGGTTTCCCTTTG     AATTCCCCATGGTATTGTTGCCGG     AATTCTTTCCGGCCCCC
```

by *EcoRI*. The restriction enzyme *NotI* is defined by

```
5'-GC'GGCCGC-3'
```

which indicates that it will find all occurrences of the string `GCGGCCGC` in the 5' to 3' direction and cut the DNA between the first `C` and the second `G`.

You may have noticed that if you form the complement of `GAATTC`, you get `CTTAAG`, which is the string spelled backwards. Similarly, the complement of `GCGGCCGC` is `CGCCGGCG`, which is also the string spelled backwards. Certain types of restriction enzymes have this *palindromic* property.

Some restriction enzymes have a cleavage site outside of the recognition site. *AceIII* is defined by

```
CAGCTCNNNNNNN'
```

The `N` matches any of `A`, `C`, `G`, or `T`. Therefore, this enzyme cuts the DNA between the 7th and 8th nucleotides after its recognition site. For example

```
CAGCTCAAATGCCAGGGGGGG
```

will be cut between the `A` and the `G`:

```
CAGCTCAAATGCCA GGGGGGG
```

In actuality, many of these restriction enzymes cleave both strands of the DNA at once, and not necessarily at the same position. We are going to simplify the problem in this assignment and assume that the DNA is single-stranded and that it is cut as described above.