

Bioinformatics: Life Sciences on Your Computer

Basics/Language of Molecular Biology

Since there are four nucleotides in DNA, we can say that the DNA language has a four-letter alphabet. There are 20 amino acids (plus two very rare ones), so protein sequences have a 22-letter alphabet. This short reading should help guide you through the basics of the process to get you started.

Summary

DNA carries genetic information (Protein does not).
DNA structure is double helical.
Sequence of four bases in DNA carries the information.
Central Dogma: DNA to RNA to protein.
Proteins are synthesized from amino-terminus to carboxy-terminus.
There are 20 amino acids that make up most proteins.

Important Definitions

DNA replication - The process by which the two strands of a DNA molecule are duplicated to generate two exact copies of the original DNA molecule.

Transcription - The process of synthesizing RNA using a DNA template. The product is a single-stranded RNA molecule that is the complement of the DNA template.

Translation - The process of synthesizing new proteins according to the directions of an mRNA. So called, because the net result is the translation of nucleic acid information into protein information.

DNA basics

There are four nucleotides that make up the sequence of DNA. Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Whenever genome sequencing is discussed, that refers to determining the order of billions these four chemicals in chromosomes. On a computer, a sequence may look like this:

```
>gi|109148525|ref|NM_000207.2| Homo sapiens insulin (INS), transcript variant 1, mRNA
AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCTTCTGCCATGGCCCTGTG
GATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAAC
CAACACCTGTGCGGCTCACACCTGGTGGAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACA
CACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGC
AGGCAGCCTGCAGCCCTTGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACC
AGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCG
CCGCCTCCTGCACCGAGAGAGATGGAATAAAGCCCTTGAACCAGCAAAA
```

This is the FASTA (pronounced "FAST-A") format for a DNA sequence. The first line starts with > to indicate a title. Then come the As, Cs, Ts and Gs, 469 of them for this human insulin gene.

Chargaff's rules

A and T are found in equal amounts in DNA and G and C are found in equal amounts. That observation helped lead to the double helix structure. James Watson & Francis Crick determined the actual double helical structure based on X-ray diffraction work, much of it done by Rosalind Franklin. A pairs with T and G pairs with C, which explains Chargaff's rules.

Discussion topic: Who was Rosalind Franklin? Did she deserve equal (or greater) billing than Watson and Crick?

Note that the DNA sequence above lists only a string of A/T/G/C. That represents only one strand. On a computer, it would be redundant to list the other (complementary) strand's sequence. For instance--the first three nucleotides are AGC. We know that the complementary strand would have TCG at that location, due to Chargaff's rules of pairing.

RNA

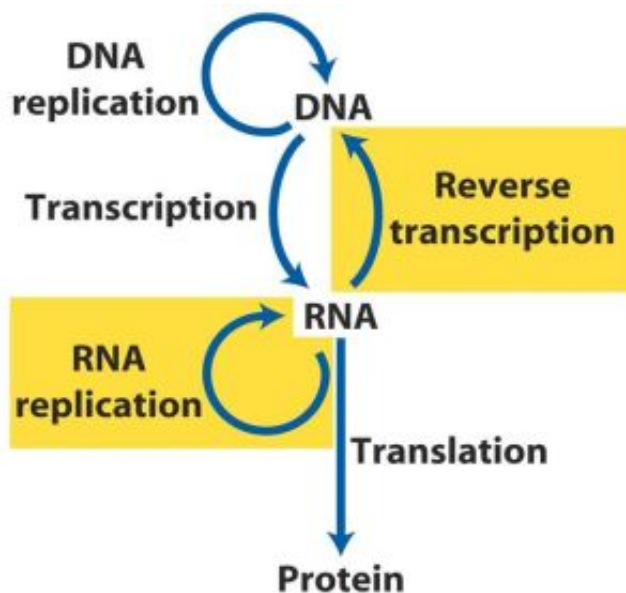
RNA is similar to DNA, but the 2' carbon has an -OH group (in DNA, it has only an -H). That oxygen makes RNA a better candidate for splicing, but that's a topic for another unit. Uracil instead of thymine is the other major difference, chemically.

Link between DNA, protein

Vernon Ingram and John Hunt's classic experiment studying hemoglobin and sickle-cell anemia, published in 1958, was the first experimental evidence leading to the notion that DNA controls amino acid sequence. In the case of sickle-cell anemia, two mutant hemoglobin genes lead to the sickle-cell phenotype, but one bad gene causes a half-sickle, half normal phenotype (which carries with it malaria resistance). Ingram and Hunt identified amino acid substitutions in the mutant protein.

Central Dogma of Biology

The central dogma written by Francis Crick in 1956 is the white part of the diagram below, which includes DNA replication, transcription and translation. The yellow boxes, including reverse transcription and RNA replication came at a later date when more was known about viruses and retrotransposons.



--image from Nelson & Cox. *Lehninger Principles of Biochemistry*

The jargon!

This is important, because this jargon is used repeatedly by biologists and bioinformaticians.

1. DNA is transcribed into RNA (transcription)
2. RNA is translated into protein (translation)

It makes some sense. RNA is a copy of DNA, but with Us instead of Ts---transcribing refers to making that copy.

The RNA to protein step requires a genetic code, so it's a different language in a sense. Translation.

Adaptor hypothesis

In 1955, Crick challenged the prevailing belief that proteins were built directly on the RNA template. The thinking was that the amino acids fit certain RNA patterns. But bases are water soluble and many amino acids are not. Crick's adaptor hypothesis suggested that specific molecules link amino acids to the template. We now know these molecules as transfer RNAs (tRNAs).

Three Major RNAs

Transfer RNA (tRNA): Transfer RNA was discovered in cell-free protein synthesis. Amino acids first attach to these specific molecules, which have short RNA chains that link a codon to its appropriate amino acid.

Ribosomal RNA (rRNA): Once thought to be template for protein synthesis, rRNA makes up 85% of cellular RNA. There are large and small subunits, each having a uniform size and base composition. The lack of variability in base composition is a strong indication that rRNA does not code for proteins (which come in large varieties). We now know the ribosome as the location of protein synthesis.

Messenger RNA (mRNA): Making up only 4% of total RNA, mRNA is the template for protein synthesis. The large variations in lengths of mRNAs correlate well with variations in protein size.

Polymerases

Polymerases are enzymes that catalyze the biosynthesis of polymers, or chains of subunits. DNA and RNA are polymers of nucleic acids. Most polymerases require a template. There are four polymerase types, depending on the nature of the template and the nature of the polymer. A DNA-directed DNA polymerase (commonly called a DNA polymerase) catalyzes the biosynthesis of DNA from a DNA template. A DNA-directed RNA polymerase (commonly called an RNA polymerase) catalyzes the biosynthesis of RNA from a DNA template (transcription). Here are the four nucleic acid polymerase types:

<i>enzyme name</i>	<i>common name</i>	<i>function</i>
DNA-directed DNA polymerase	DNA polymerase	Replication of DNA
DNA-directed RNA polymerase	RNA polymerase	Transcription
RNA-directed DNA polymerase	reverse transcriptase	retroviral replication (DNA from RNA genome)
RNA-directed RNA polymerase	--	viral replication, synthesis of antisense RNA

RNA polymerase

The DNA-directed RNA polymerase (RNA polymerase) catalyzes the synthesis of RNA from a DNA template (transcription). RNA polymerases are used to synthesize mRNA, rRNA, tRNA. The DNA has two strands with respect to a molecule of RNA: the template and the nontemplate strand.

Template and nontemplate strand

Used in transcription, the template strand has a sequence complementary to the mRNA sequence. The nontemplate strand, which is also called the "coding strand," is also complementary to template strand. It is called the coding strand because its sequence matches the sequence of mRNA, except that it has Ts instead of Us.

(5') CGCTATAGCGTTT (3')	DNA nontemplate (coding) strand
(3') GCGATATCGCAA (5')	DNA template strand
(5') CGCUAUAGCGUUU (3')	RNA transcript

-image from Nelson & Cox. *Lehninger Principles of Biochemistry*

Translation

RNA binds a series of ribosomes. The series can be called a polyribosome. Transfer RNA brings in amino acids. With roughly one tRNA per ribosome, peptide bonds are formed here.

Cracking genetic code

Three nucleotides code for one amino acid in a code that was elucidated in 1966. Sydney Brenner & Crick in 1961 used synthetic RNAs used to begin to crack the code. UUUUUUUUUU only codes for phenylalanine, suggesting that UUU is the codon for phenylalanine. The stop codons were the easiest to find.

		Second letter				Third letter
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	

--image from geek.com

Note that the triplet is called a codon. There are 64 triplets but only 20 amino acids, so some amino acids are coded for by more than one triplet. Some mutations, or changes in the DNA sequence, do not have any effect on protein sequence.

Direction of protein synthesis

From pulse labeling experiments, it was found that the most recently added amino acids during protein synthesis were found at the carboxy-terminus (C-terminus) of the protein. Conclusion: protein synthesis begins at the amino-terminus (N-terminus) and proceeds to the carboxy-terminus (C-terminus).

Therefore, protein sequences are always written with beginning with the N-terminus and ending with the C-terminus.