The Central Dogma and underlying molecular biology

This topic is reviewed in BI229F *not* because it is *per se* part of the course, but because you need to have a reasonable understanding of the most common ways in which information is encoded and used in order for the things that *are* part of the course to make sense.

The exam will have some questions related to this lecture that are designed to test your general understanding but not your detailed knowledge.

This document has been roughly divided up into sections related to the slides given in the lecture and clarifies what I expect you to understand and for which there may be questions in the exam.

What is DNA and RNA, what are their similarities and differences

General

1. Both are polymers of nucleotides; DNA molecules are generally longer than RNA molecules.
2. DNA is *generally* a form of *read only* memory that is passed on unchanged from generation to generation[1]. The *sequence* of RNA is generally copied from a DNA sequence through *transcription*[2].
3. In general, DNA is double stranded and RNA is single stranded. But this *isn't* part of the definition of DNA and RNA.
4. Base pairing giving rise to double helices (whether RNA or DNA) in an anti-parallel manner; that is, the orientation of the two single stranded molecules are aligned in opposite orientations. We refer to two sequences as the forward and reverse strands, but which is forward and which is reverse is often arbitrary. The two strands are *complementary* to each other. To obtain the complementary sequence you must reverse the sequence of complementary nucleotides (i.e. *reverse complement* the sequence).
5. You may sometimes see sequences specified with 5' and 3' ends (eg. `5' ATCAGA 3'`). That is just to emphasise the orientation of the strand. In general, all sequences are specified in the 5' to 3' direction, and the 5' and 3' notation is almost only used in educational material[3].

[1] We often refer to the DNA sequence as not changing; this is of course not true. The DNA sequence will change both through random mutation and through processes like meiotic recombination and during the generation of specific antibodies (including both recombination events and hypermutation).

[2] We often talk of DNA being copied into RNA, but it is only the *sequence* of the DNA that is copied; i.e. the information that is stored in the DNA is transmitted to the RNA molecule.

[3] It may also be used when specifiying the sequence of short DNA sequences (oligonucleotides) that are used as probes against RNA or DNA sequences. In such cases the sequence is defined against the forward strand and so is the reverse complement of the target sequence and the notation is sometimes used to avoid confusion; but it is not essential. 5' and 3' labels are also

6. Single stranded nucleic acid sequences will form secondary structures unless otherwise prevented. Double stranded sequences have a much more stable structure that is only minimally affected by the sequence; this means that double helices can contain virtually any sequence whereas single stranded sequences will be constrained by their tendency to fold up[4].
7. Nucleic acid sequences are always synthesised from the 5' end with new nucleotides being added at the 3'-OH (hydroxyl) group (often just referred to as the "three-prime end").
8. In bioinformatics we are usually only concerned with the sequence of DNA molecules and this is always given in the 5' to 3' direction.

Nucleotides

1. There are four identities: A, C, G and T (DNA) or U (RNA)[5].
2. The identity is defined by the base of the nucleotide, not by the backbone, and it is the sequence of bases that encodes information.
3. The difference between RNA and DNA nucleotides lies in the ribose-phosphate backbone (DNA nucleotides have one less oxygen molecule, deoxyribonucleotide vs ribonucleotide). This stabilises DNA molecules.
4. The nucleotides have polarity or direction with a 5' and a 3' end; 5' ends are joined to 3' ends in the polymer.
5. Pairs of bases can base-pair through hydrogen bonds in an antiparallel manner (opposing direction). A forms pairs with T or U (RNA) and C pairs with G.
6. Base-pairing (and thus double helix formation) can occur between DNA-DNA, DNA-RNA and RNA-RNA sequence pairs.
7. A-T base pairs are somewhat less stable than G-C pairs (due to having 2 rather than 3 hydrogen bonds).

Copying, packaging and regulation

You do not need to know many details of how DNA is copied or packaged, but it probably helps to have some idea about it.

You should be aware of the fact that chromosomes (the individual DNA molecules that form the genome) are packed around protein molecules (histones[6]). Histones can be modified by a number of different mechanisms and this modifications are related to the regulation of gene activity through the interaction with transcrip-

used in figures giving the structure of nucleic acid sequences and in figures generally.

[4]Note that there are a multitude of proteins that are used to suppress the tendency of single stranded nucleic acids to fold up and that mRNA molecules are not naked strands of RNA floating around in the cell but are bound by large numbers of proteins.

[5]We usually consider U and T to be equivalent to each other with the only difference being that U is incorporated into RNA molecules and T in DNA molecules. But not surprisingly, they *do* have somewhat different properties. Note that there *are* more than four types of ribonucleotides (eg. Inosine), but that we will not care about that here.

[6]In sperm cells the DNA is much more packaged around proteins called protamines instead of the usual histamine molecules. You do not need to know this.

tion factors (proteins that regulate gene expression) and DNA methylation (a covalent modification of the DNA itself).

It is often stated that the double-strandedness of DNA facilitates its being copied; this is probably often repeated because of a famous sentence in the Watson and Crick paper. However, single stranded DNA and RNA genomes do exist, so double-strandedness clearly isn't a requirement for genomes to be copied.

Genes and transcription

No single definition of a gene exists, but we usually mean, 'a region that has function, part of which is transcribed into RNA and part of which affects the regulation of transcription'.

In bioinformatics we are often concerned with measurements of the expression of large numbers of genes in different samples. Different cell types express different sets of genes and gene expression can be changed by environmental changes (eg. blood sugar levels).

Prokaryotic genes may encode several proteins transcribed in a single polycistronic transcript. Eukaryotic genes usually (depends on organism) contain introns that have to be removed before the RNA molecule can be translated. The part of the sequence that remains are the exons. The process of removing the exons is referred to as splicing. Genes can encode several different transcripts which contain different sets of exons (alternative splicing).

Note that non-coding transcripts can also contain introns and be spliced.

Genes are transcribed from a region referred to as a 'promoter'. The sequence at the promoter and other sequences around the gene (can be both upstream and downstream as well as in intronic regions) affect the cell types and circumstances under which the gene is transcribed (active). The non-promoter regulatory sequences are called 'cis-acting' regions; they can be distant from the gene and can be shared between genes.

You should have a basic idea of the difference in eukaryotic and prokaryotic gene transcription (splicing, cistrons, poly-adenylation).

Proteins and the genetic code

1. Proteins are also polymer molecules; but made up by sequences of 20 different amino acids.
2. They have both structural and catalytic functions (which includes the transmission of information and transcriptional regulation).
3. The differences in the chemical properties of the amino acids allows proteins to form almost any shape.
4. The sequence of amino acids in a protein determines it's structure and properties.

5. Protein sequences are encoded in the genome (usually DNA) as nucleotide triplets called codons. There are 61 codons encoding 20 amino acids and 3 codons that indicate 'stop of translation'.
6. There are a number of 'genetic codes', but they have only small number of differences.

The central dogma and some exceptions.

As usual, more than one definition of the central dogma exists. The most common one states that DNA contains sequences that encode protein sequences; these sequences are copied into RNA molecules through transcription and the sequences in the RNA molecules are converted to amino acid sequences (proteins) during translation at ribosomes.

This commonly stated central dogma has a number of exceptions that you should be aware of:

1. Functional RNA molecules whose sequences are not translated into proteins.
2. RNA viruses, whose genomes are composed of RNA sequences not DNA sequences.
3. Retroviruses; viruses with RNA genomes whose sequences can be copied into DNA molecules which can then integrate into a DNA genome.
4. Retrotransposons; DNA sequences that are transcribed into RNA, then reverse transcribed into DNA molecules that can integrate into the genome. These can be described as endogenous retroviruses.

A more specific definition states simply that information can flow from DNA to proteins via RNA molecules but that information present in protein sequences can not flow back to RNA or DNA molecules. That is, we know of no mechanism that can translate an amino acid sequence to a nucleotide sequence.

Important Information

You need to understand the polarity (orientation) of nucleic acid sequences and how this affects the conversion to amino acid sequences. To some extent this is dependant on convention. For example, for a DNA sequence containing a gene we will almost always be given the forward sequence of the gene. That means that transcription starts from somewhere close to the beginning of the sequence and ends somewhere close to the end. Usually you will not know exactly where. The sequence of the RNA molecule that is transcribed is the same as the DNA sequence given (it is *not* the complement), except with T's replaced with U's.

If you are given a DNA sequence with no further information you can assume it is double stranded; this means that either strand could encode protein sequence and that there are 6 ways in which it could be translated (via an mRNA molecule) to a protein sequence.

Conversely if you are given an RNA sequence, you may infer that it is single stranded and hence that it only has 3 reading frames.

You should be able to give all possible translations of a given DNA or RNA molecule (given the genetic code) and also be able to argue about why codons have a length of three and the implication this has on the effect of mutations in coding regions.